Strategic Equilibrium*

Eric van Damme

January 1994
Revised 2000

Abstract

An outcome in a noncooperative game is said to be self-enforcing, or a strategic equilibrium, if, whenever it is recommended to the players, no player has an incentive to deviate from it. This paper gives an overview of the concepts that have been proposed as formalizations of this requirement and of the properties and the applications of these concepts. In particular the paper discusses Nash equilibrium, together with its main coarsenings (correlated equilibrium, rationalizibility) and its main refinements (sequential, perfect, proper, persistent and stable equilibria). There is also an extensive discussion on equilibrium selection.

*This paper was written in 1994, and no attempt has been made to provide a survey of the developments since then. The author thanks two anonymous referees and the editors for their comments.

1 Introduction

It has been said that "the basic task of game theory is to tell us what strategies rational players will follow and what expectations they can rationally entertain about other rational players' strategies" (Harsanyi and Selten (1988, p. 342)). To construct such a theory of rational behavior for interactive decision situations, game theorists proceed in an indirect, roundabout way, as suggested in Von Neumann and Morgenstern (1944, §17.3). The analyst assumes that a satisfactory theory of rational behavior exists and tries to deduce which outcomes are consistent with such a theory. A fundamental requirement is that the theory should not be self-defeating, i.e. players who know the theory should have no incentive to deviate from the behavior that the theory recommends. For noncooperative games, i.e. games in which there is no external mechanism available for the enforcement of agreements or commitments, this requirement implies that the recommendation has to be self-enforcing. Hence, if the participants act independently and if the theory recommends a unique strategy for each player, the profile of recommendations has to be a Nash equilibrium: The strategy that is assigned to a player must be optimal for this player when the other players follow the strategies that are assigned to them. As Nash writes

> "By using the principles that a rational prediction should be unique, that the players should be able to make use of it, and that such knowledge on the part of each player of what to expect the others to do should not lead him to act out of conformity with the prediction, one is led to the concept" (Nash (1950a)).

Hence, a satisfactory normative theory that advises people how to play games necessarily must prescribe a Nash equilibrium in each game. Consequently, one wants to know whether Nash equilibria exist and what properties they have. These questions are addressed in the next section of this paper. In that section we also discuss the concept of rationalizability, which imposes necessary requirements for a satisfactory set-valued theory of rationality. A second immediate question is whether a satisfactory theory can prescribe just any Nash equilibrium, i.e. whether all Nash equilibria are self-enforcing.

Simple examples of extensive form games have shown that the answer to this question is no: Some equilibria are sustained only by incredible threats and, hence, are not viable as the expectation that a rational player will carry out an irrational (nonmaximizing) action is irrational. This observation has stimulated the search for more refined equilibrium notions that aim to formalize additional necessary conditions for self-enforcingness. A major part of this paper is devoted to a survey of the most important of these so-called refinements of the Nash equilibrium concept. (See Chapter 62 in this Handbook for a general critique on this refinement program.)

In Section 3 the emphasis is on extensive form solution concepts that aim to capture the idea of backward induction, i.e. the idea that rational players should be assumed to be forward-looking and to be motivated to reach their goals in the future, no matter what happened in the past. The concepts of subgame perfect, sequential, perfect and proper equilibria that are discussed in Section 3 can all be viewed as formalizations of this basic idea. Backward induction, however, is only one aspect of self-enforcingness, and it turns out that it is not sufficient to guarantee the latter. Therefore, in Section 4 we turn to another aspect of self-enforcingness, that of forward induction. We will discuss stability concepts that aim at formalizing this idea, i.e. that actions taken by rational actors in the past should be interpreted, whenever possible, as being part of a grand plan that is globally optimal. As these concepts are related to the notion of persistent equilibrium, we will have an opportunity to discuss this latter concept as well. Furthermore, as these ideas are most easily discussed in the normal form of the game, we take a normal-form perspective in Section 4. As the concepts discussed in this section are set-valued solution concepts, we will also discuss the extent to which set-valuedness contradicts the uniqueness of the rational prediction as postulated by Nash in the above quotation.

The fact that many games have multiple equilibria poses a serious problem for the "theory" rationale of Nash equilibrium discussed above. It seems that, for Nash's argument to make sense, the theory has to select a unique equilibrium in each game. However, how can a rational prediction be unique if the game has multiple equilibria? How can one rationally select an equilibrium? A general approach to this latter problem has

been proposed in Harsanyi and Selten (1988), and Section 5 is devoted to an overview of that theory as well as a more detailed discussion of some of its main elements, such as the tracing procedure and the notion of risk-dominance. We also discuss some related theories of equilibrium selection in that section and show that the various elements of self-enforcingness that are identified in the various sections may easily be in conflict; hence, the search for a universal solution concept for non-cooperative games may continue in the future.

I conclude this introduction with some remarks concerning the (limited) scope of this chapter. As the Handbook contains an entire chapter on the conceptual foundations of strategic equilibrium (Chapter 42 of this Handbook), there are few remarks on this topic in the present chapter. I do not discuss the epistemic conditions needed to justify Nash equilibrium (see Aumann and Brandenburger (1995)), nor how an equilibrium can be reached. I'll focus on the formal definitions and mathematical properties of the concepts. Throughout, attention will be restricted to finite games, i.e. games in which the number of players as well as the action set of each of these players is finite. It should also be stressed that several other completely different rationales have been advanced for Nash equilibria, and that these are not discussed at all in this chapter. Nash (1950a) already discussed the "mass-action" interpretation of equilibria, i.e. that equilibria can result when the game is repeatedly played by myopic players who learn over time. I refer to Fudenberg and Levine (1998), and the papers cited therein for a discussion of the contexts in which learning processes can be expected to converge to Nash equilibria. Maynard Smith and Price (1973) showed that Nash equilibria can result as outcomes of evolutionary processes that wipe out less fit strategies through time. I refer to Hammerstein and Selten (1994) and Van Damme (1994) for a discussion of the role of Nash equilibrium in the biological branch of game theory, and to Samuelson (1997), Vega-Redondo (1996) and Weibull (1995) for more general discussions on evolutionary processes in games.

2 Nash Equilibria in normal form games

2.1 Generalities

A (finite) *game in normal form* is a tuple $g = (A, u)$ where $A = A_1 \times ... \times A_I$ is a Cartesian product of finite sets and $u = (u_1, ..., u_I)$ is an I-tuple of functions $u_i : A \to \mathsf{R}$. The set $I = \{1, ..., I\}$ is the set of players, A_i is the set of pure strategies of player i and u_i is this player's payoff function. Such a game is played as follows: Simultaneously and independently players choose strategies; if the combination $a \in A$ results, then each player i receives $u_i(a)$. A *mixed strategy* of player i is a probability distribution s_i on A_i and we write S_i for the set of such mixed strategies, hence

$$S_i = \{s_i : A_i \to \mathsf{R}_+, \quad \sum_{a_i \in A_i} s_i(a_i) = 1\}. \tag{2.1}$$

(Generally, if C is any finite set, $\Delta(C)$ denotes the set of probability distributions on C, hence, $S_i = \Delta(A_i)$). A mixed strategy may be interpreted as an act of deliberate randomization of player i or as a probability assessment of some player $j \neq i$ about how i is going to play. We return to these different interpretations below. We identify $a_i \in A_i$ with the mixed strategy that assigns probability 1 to a_i. We will write S for the set of mixed strategy profiles, $S = S_1 \times ... \times S_I$, with s denoting a generic element of S. Note that when strategies are interpreted as beliefs, taking strategy profiles as the primitive concept entails the implicit assumption that any two opponents j, k of player i have a common belief s_i about which pure action i will take. Alternatively, interpreting s as a profile of deliberate acts of randomization, the expected payoff to i when $s \in S$ is played, is written $u_i(s)$, hence

$$u_i(s) = \sum_{a \in A} \prod_{j \in I} s_j(a_j) u_i(a). \tag{2.2}$$

If $s \in S$ and $s'_i \in S_i$, then $s \backslash s'_i$ denotes the strategy profile in which each $j \neq i$ plays s_j while i plays s'_i. Occasionally we also write $s \backslash s'_i = (s_{-i}, s'_i)$, hence, s_{-i} denotes the strategy vector used by the opponents of player i. We also write $S_{-i} = \prod_{j \neq i} S_j$ and

$A_{-i} = \prod_{j \neq i} A_j$. We say that s_i' is a *best reply* against s in g if

$$u_i(s \backslash s_i') = \max_{s_i'' \in S_i} u_i(s \backslash s_i'') \tag{2.3}$$

and the set of all such best replies is denoted as $\mathcal{B}_i(s)$. Obviously, $\mathcal{B}_i(s)$ only depends on s_{-i}, hence, we can also view $\mathcal{B}_i$ as a correspondence from S_{-i} to S_i. If we write $B_i(s)$ for the set of pure best replies against s, hence $B_i(s) = \mathcal{B}_i(s) \cap A_i$, then obviously $\mathcal{B}_i(s)$ is the convex hull of $B_i(s)$. We write $\mathcal{B}(s) = \mathcal{B}_1(s) \times \ ... \ \times \mathcal{B}_I(s)$ and refer to $\mathcal{B} : S \to S$ as the *best-reply correspondence* associated with g. The pure best reply correspondence is denoted by B, hence $B = \mathcal{B} \cap A$.

2.2 Self-enforcing theories of rationality

We now turn to solution concepts that try to capture the idea of a theory of rational behavior being self-enforcing. We assume that it is common knowledge that players are rational in the Bayesian sense, i.e. whenever a player faces uncertainty, he constructs subjective beliefs representing that uncertainty and chooses an action that maximizes his subjective expected payoffs. We proceed in the indirect way outlined in Von Neumann and Morgenstern (1944, §17.3). We assume that a self-enforcing theory of rationality exists and investigate its consequences, i.e. we try to determine the theory from its necessary implications. The first idea for a solution of the game g is a definite strategy recommendation for each player, i.e. some $a \in A$. Already in simple examples like matching pennies, however, no such simple theory can be self-enforcing: There is no $a \in A$ that satisfies $a \in B(a)$, hence, there is always at least one player who has an incentive to deviate from the strategy that the theory recommends for him. Hence, a general theory of rationality, if one exists, must be more complicated.

Let us now investigate the possibilities for a theory that may recommend more than one action for each player. Let $C_i \subset A_i$ be the nonempty set of actions that the theory recommends for player i in the game g and assume that the theory, i.e. the set $C = X_i C_i$, is common knowledge among the players. If $|C_j| > 1$, then player i faces uncertainty about player j's action, hence, he will have beliefs $s_j^i \in S_j$ about what j will do. Assuming beliefs associated with different opponents to be independent, we

can represent player i's beliefs by a mixed strategy vector $s^i \in S_{-i}$. (Below we also discuss the case of correlated beliefs; a referee remarked that he considered that to be the more relevant case.) The crucial question now is which beliefs can player i rationally entertain about an opponent j. If the theory C is self-enforcing, then no player j has an incentive to choose an action that is not recommended, hence, player i should assign zero possibility to any $a_j \in A_j \backslash C_j$. Writing $C_j(s_j)$ for the support of $s_j \in S_j$,

$$C_j(s_j) = \{a_j \in A_j : s_j(a_j) > 0\}, \tag{2.4}$$

we can write this requirement as

$$C_j(s_j^i) \subset C_j \text{ for all } i, j. \tag{2.5}$$

The remaining question is whether all beliefs s_j^i satisfying (2.5) should be allowed, i.e. whether i's beliefs about j can be represented by the set $\Delta(C_j)$. One might argue yes: If the opponents of j had an argument to exclude some $a_j \in C_j$, our theory would not be very convincing; the players would have a better theory available (simply replace C_j by $C_j \backslash \{a_j\}$). Hence, let us insist that all beliefs s_j^i satisfying (2.5) are allowed. Being Bayesian rational, player i will choose a best response against his beliefs s^i. His opponents, although not necessarily knowing his beliefs, know that he behaves in this way, hence, they know that he will choose an action in the set

$$B_i(C) = \bigcup \{B_i(s^i) : s_j^i \in \Delta(C_j) \text{ for all } j\}. \tag{2.6}$$

Write $B(C) = X_i B_i(C)$. A necessary requirement for C to be self-enforcing now is that

$$C \subset B(C). \tag{2.7}$$

For, if there exists some $i \in I$ and some $a_i \in A_i$ with $a_i \in C_i \backslash B_i(C)$, then the opponents know that player i will not play a_i, but then they should assign probability zero to a_i, contradicting the assumption made just below (2.5). Write 2^A for the collection of subsets of A. Obviously, 2^A is a finite, complete lattice and the mapping $B : 2^A \to 2^A$ (defined by (2.6) and $B(\emptyset) = \emptyset$) is monotonic. Hence, it follows from Tarski's fixed point theorem (Tarski (1955)), or by direct verification that

(i) there exists a nonempty set C satisfying (2.7),

(ii) the set of all sets satisfying (2.7) is again a complete lattice, and

(iii) the union of all sets C satisfying (2.7), to be denoted R, is a fixed point of B, i.e. $R = B(R)$, hence, R is the largest fixed point.

The set R is known as the set of pure *rationalizable strategy profiles* in g (Bernheim (1984), Pearce (1984)). It follows by the above arguments that any self-enforcing set-valued theory of rationality has to be a subset of R and that R itself is such a theory. The reader can also easily check that R can be found by repeatedly eliminating the non-best responses from g, hence

$$\text{if } C^0 = A \text{ and } C^{t+1} = B(C^t), \text{ then } R = \bigcap_t C^t. \tag{2.8}$$

It is tempting to argue that, for C to be self-enforcing, it is not only necessary that (2.7) holds, but also that conversely

$$B(C) \subset C; \tag{2.9}$$

hence, that C actually must be a fixed point of B. The argument would be that, if (2.9) did not hold and if $a_i \in B_i(C) \backslash C_i$, player i could conceivably play a_i, hence, his opponents should assign positive probability to a_i. This argument, however, relies on the assumption that a rational player can play any best response. Since not all best responses might be equally good (some might be dominated, inadmissible, inferior or non-robust (terms that are defined below)), it is not completely convincing. We note that sets with the property (2.9) have been introduced in Basu and Weibull (1991) under the name of *curb sets*. (Curb is mnemonic for closed under rational behavior.) The set of all sets satisfying (2.9) is a complete lattice, i.e. there are minimal nonempty elements and such minimal elements are fixed points. (Fixed points are called tight curb sets in Basu and Weibull (1991).) We will encounter this concept again in Section 4.

Above we allowed two different opponents i and k to have different beliefs about player j, hence $s_j^i \neq s_j^k$. In such situations one should actually discuss the beliefs that i has about k's beliefs. To avoid discussing such higher-order beliefs, let us assume that players' beliefs are summarized by one strategy vector $s \in S$, hence we are discussing

a theory that recommends a unique mixed strategy vector. For such a theory s to be self-enforcing, we obtain, arguing exactly as above, as a necessary requirement

$$C(s) \subset B(s) \tag{2.10}$$

where $C(s) = X_i C_i(s_i)$, hence, each player believes that each opponent will play a best response against his beliefs. A condition equivalent to (2.10) is

$$s \in \mathcal{B}(s) \tag{2.11}$$

or

$$u_i(s) = \max_{s_i' \in S_i} u_i(s \backslash s_i') \quad \text{for all} \;\; i \in I. \tag{2.12}$$

A strategy vector s satisfying these conditions is called a *Nash equilibrium* (Nash (1950b, 1951)). A standard application of Kakutani's fixed point theorem yields:

Theorem 1 *(Nash (1950b, 1951)). Every (finite) normal form game has at least one Nash equilibrium.*

We note that Nash (1951) provides an elegant proof that relies directly on Brouwer's fixed point theorem. We have already seen that some games only admit equilibria in mixed strategies. Dresher (1970) has computed that a large game with randomly drawn payoffs has a pure equilibrium with probability $1 - 1/e$. More recently, Stanford (1995) has derived a formula for the probability that a randomly selected game has exactly k pure equilibria. Gul et al. (1993) have shown that, for generic games , if there are $k \geq 1$ pure equilibria, then the number of mixed equilibria is at least $2k - 1$, a result to which we return below. An important class of games that admit pure equilibria are *potential games* (Monderer and Shapley (1996)). A function $P : A \to \mathsf{R}$ is said to be an *ordinal potential* of $g = \langle A, u \rangle$ if for every $a \in A, i \in I$ and $a_i' \in A_i$

$$u_i(a) - u_i(a \backslash a_i') > 0 \;\; \text{iff} \;\; P(a) - P(a \backslash a_i') > 0. \tag{2.13}$$

Hence, if (2.13) holds, then g is ordinally equivalent to a game with common payoffs and any maximizer of the potential P is a pure equilibrium of g. Consequently, a game g that has an ordinal potential, has a pure equilibrium. Note that g may have pure

equilibria that do not maximize P and that there may be mixed equilibria as well. The function P is said to be an *exact potential* for g if

$$u_i(a) - u_i(a\backslash a_i') = P(a) - P(a\backslash a_i') \tag{2.14}$$

and Monderer and Shapley (1996) show that such an exact potential, when it exists, is unique up to an additive constant. Hence, the set of all maximizers of the potential is a well-defined refinement. Neyman (1997) shows that if the multilinear extension of P from A to S (as in (2.2)) is concave and continuously differentiable, every equilibrium of g is pure and is a maximizer of the potential. Another class of games, with important applications in economics, that admit pure strategy equilibria are *games with strategic complementaries* (Topkis (1979), Vives (1990), Milgrom and Roberts (1990, 1991), Milgrom and Shannon (1994)). These are games in which each A_i can be ordered so that it forms a complete lattice and in which each player's best-response correspondence is monotonically nondecreasing in the opponents' strategy combination. The latter is guaranteed if each u_i is supermodular in a_i (i.e. $u_i(a_i, a_{-i}) + u_i(a_i', a_{-i}) \leq u_i(a_i \wedge a_i', a_{-i}) + u_i(a_i \vee a_i', a_{-i})$) and has increasing differences in (a_i', a_{-i}) (i.e. if $a_{-i} \geq a_{-i}'$, then $u_i(a_i, a_{-i}) - u_i(a_i, a_{-i}')$ is increasing in a_i). Topkis (1979) shows that such a game has at least one pure equilibrium and that there exists a largest and a smallest equilibrium, $\bar{a}$ and $\underline{a}$ respectively. Milgrom and Roberts (1990, 1991) show that $\bar{a}_i$ (resp. $\underline{a}_i$) is the largest (resp. smallest) serially undominated action of each player i, hence, by iterative elimination of strictly dominated strategies, the game can be reduced to the interval $[\underline{a}, \bar{a}]$. It follows that, if a game with strategic complementarities has a unique equilibrium, it is dominance-solvable, hence, that only the unique equilibrium strategies are rationalizable.

An equilibrium s^* is called *strict* if it is the unique best reply against itself, hence $\{s^*\} = B(s^*)$. Obviously, strict equilibria are necessarily in pure strategies, consequently they need not exist. An equilibrium s^* is called *quasi-strict* if all pure best replies are chosen with positive probability in s^*, that is, if $a_i \in B_i(s^*)$, then $s_i^*(a_i) > 0$. Also, quasi-strict equilibria need not exist: Van Damme (1987a, p. 56) gives a 3-player example. Norde (1999) has shown, however, that quasi-strict equilibria do exist in 2-person games.

An axiomatization of the Nash concept, using the notion of *consistency*, has been provided in Peleg and Tijs (1996). Given a name g, a strategy profile s and a coalition of players C, define the reduced game $g^{C,s}$ as the game that results from g if the players in $I\backslash C$ are committed to play strategies as prescribed by s. A family of games Γ is called closed if all possible reduced games, of games in Γ, again belong to Γ. A solution concept on Γ is a map φ that associates to each g in Γ a non-empty set of strategy profiles in g. φ is said to satisfy one-person rationality (OPR) if in every one-person game it selects all payoff maximizing actions. On a closed set of games Γ, φ is said to be consistent (CONS) if, for every g in Γ and s and C: if $s \in \varphi(g)$, then $s_C \in \varphi(g^{C,s})$, in other words, if some players are committed to play a solution, the remaining players find that the solution prescribed to them is a solution for their reduced game. Finally, a solution concept φ on a closed set Γ is said to satisfy converse consistency (COCONS) if, whenever s is such that $s_C \in \varphi(g^{C,s})$ for all $C \neq \phi$, then also $s \in \varphi(g)$; in other words, if the profile is a solution in all reduced games, then it is also a solution in the overall game. Peleg and Tijs (1996, Theorem 2.12) show that, on any closed family of games, the Nash equilibrium correspondence is characterized by the axioms OPR, CONS and COCONS.

Next, let us briefly turn to the assumption that strategy sets are finite. We note, first of all, that Theorem 1 can be extended to games in which the strategy sets A_i are nonempty, compact subsets of some finite-dimensional Euclidean space and the payoff functions u_i are continuous (Glicksberg (1952)). If, in addition, A_i is convex and u_i is quasi-concave in a_i, there exists a pure equilibrium. Existence theorems for discontinuous games have been given in Dasgupta and Maskin (1986) and Simon and Zame (1990). In the latter paper it is pointed out that discontinuities typically arise from indeterminacies in the underlying (economic) problem and that these may be resolved by formulating an endogenous sharing rule. In this paper, emphasis will be on finite games. All games will be assumed finite, unless explicitly stated otherwise.

To conclude this subsection, we briefly return to the independence assumption that underlies the above discussion, i.e. the assumption that player i represents his uncertainty about his opponents by a mixed strategy vector $s^i \in S_{-i}$. A similar development is pos-

sible if we allow for correlation. In that case, (2.8) will be replaced by the procedure of iterative elimination of strictly dominated strategies, and the analogous concept to (2.9) is that of formations (Harsanyi and Selten (1988), see also Section 5). The concept that corresponds to the parallel version of (2.12) is that of correlated equilibrium, Aumann (1974). Formally, if σ is a correlated strategy profile (i.e. σ is a probability distribution on A, $\sigma \in \Delta(A)$), then σ is a *correlated equilibrium* if for each player i and each $a_i \in A_i$

$$\text{if } \sigma_i(a_i) > 0 \text{ then } \sum_{a_{-i}} \sigma_{-i}(a_{-i}|a_i) u_i(a_{-i}, a_i) \geq \sum_{a_{-i}} \sigma_{-i}(a_{-i}|a_i) u_i(a_{-i}, a_i') \text{ for all } a_i' \in A_i$$

where $\sigma_i(a_i)$ denotes the marginal probability of a_i and where $\sigma_{-i}(a_{-i}|a_i)$ is the conditional probability of a_{-i} given a_i. One interpretation is as follows. Assume that an impartial mediator (a person or machine through which the players communicate) selects an outcome (a recommendation) $a \in A$ according to σ and then informs each player i privately about this player's personal recommendation a_i. If the above conditions hold, then, assuming that the opponents will always follow their recommendations, no player has any incentive to deviate from his recommendation, no matter what σ may recommend to him, hence, the recommendation σ is self-enforcing. Note that correlated equilibrium allows for private communication between the mediator and each player i: After hearing his recommendation a_i, player i does not necessarily know what action has been recommended to j, and two players i and k may have different posterior beliefs about what j will do. Aumann (1974) shows that a correlated equilibrium is nothing but a Nash equilibrium of an extended game in which the possibilities for communicating and correlating have been explicitly modeled, so in a certain sense there is nothing new here, but, of course, working with a reduced form solution concept may have its advantages. More importantly, Aumann (1987a) argues that correlated beliefs arise naturally and he shows that, if it is common knowledge that each player is rational (in the Bayesian sense) and if players analyse the game by using a common prior, then the resulting distribution over outcomes must be a correlated equilibrium. Obviously, each Nash equilibrium is a correlated equilibrium, so that existence is guaranteed. An elementary proof of existence, which uses the fact that the set of correlated equilibria is a polyhedral set, has been given in Hart and Schmeidler (1989). Moulin and Vial

(1978) gives an example of a correlated equilibrium with a payoff that is outside the convex hull of the Nash equilibrium payoffs, thus showing that players may benefit from communication with the mediator not being public. Myerson (1986) shows that, in extensive games, the timing of communication becomes of utmost importance. For more extensive discussion on communication and correlation in games, we refer to Myerson's chapter 24 in this Handbook.

2.3 Structure, regularity and generic finiteness

For a game g we write $E(g)$ for the set of its Nash equilibria. It follows from (2.10) that $E(g)$ can be described by a finite number of polynomial inequalities, hence, $E(g)$ is a semi-algebraic set. Consequently, $E(g)$ has a finite triangulation, hence

Theorem 2 *(Kohlberg and Mertens (1986, Proposition 1)). The set of Nash equilibria of a game consists of finitely many connected components.*

Two equilibria s, s' of g are said to be *interchangeable* if, for each $i \in I$, also $s\backslash s'_i$ and $s'\backslash s_i$ are equilibria of g. Nash (1951) defined a *subsolution* as a maximal set of interchangeable equilibria and he called a game solvable if all its equilibria are interchangeable. Nash proved that each subsolution is a closed and convex set, in fact, that it is a product of polyhedral sets. Subsolutions need not be disjoint and a game may have uncountably many subsolutions (Chin et al. (1974)). In the 2-person case, however, there are only finitely many subsolutions (Jansen (1981)). A special class of solvable games is the 2-person *zero-sum games*, i.e. $u_1 + u_2 = 0$. For such games, all equilibria yield the same payoff, the so-called value of the game, and a strategy is an equilibrium strategy if and only if it is a minmax strategy. The reader is referred to chapter 20 in this Handbook for a more extensive discussion of zero-sum 2-person games.

Let us now take a global perspective. Write $\Gamma = \Gamma_A$ for the set of all normal form games g with strategy space $A = A_1 \times ... \times A_I$. Obviously, $\Gamma = \mathsf{R}^{I \times A}$, a finite-dimensional linear space. Write E for the graph of the equilibrium correspondence, hence, $E = \{(g, s) \in \Gamma \times S : s \in E(g)\}$. Kohlberg and Mertens have shown that this graph E is itself a relatively simple object as it is homeomorphic to the space of games Γ.

Kohlberg and Mertens show that the graph E (when compactified by adding a point ∞) looks like a deformation of a rubber sphere around the (similarly compactified) sphere of games. Hence, the graph is "simple", it just has folds, there are no holes, gaps or knots. Formally

Theorem 3 *(Kohlberg and Mertens (1986, Theorem 1)). Let π be the projection from E to Γ. Then there exists a homeomorphism φ from Γ to E such that $\pi \circ \varphi$ is homotopic to the identity on Γ under a homotopy that extends from Γ to its one-point compactification $\bar{\Gamma}$.*

Kohlberg and Mertens use Theorem 3 to show that each game has at least one component of equilibria that does not vanish entirely when the payoffs of the game are slightly perturbed, a result that we will further discuss in Section 4. We now move on to show that the graph E is really simple as generically (i.e. except on a closed set of games with measure zero) the equilibrium correspondence consists of a finite (odd) number of differentiable functions. We proceed in the spirit of Harsanyi (1973a), but follow the more elegant elaboration of Ritzberger (1994). At the end of the subsection, we briefly discuss some related recent work that provides a more general perspective.
Obviously, if s is a Nash equilibrium of g, then s is a solution to the following system of equations

$$s_i(a_i)[u_i(s\backslash a_i) - u_i(s)] = 0 \qquad \text{all } i \in I, a_i \in A_i. \tag{2.15}$$

(The system (2.15) also admits solutions that are not equilibria - for example, any pure strategy vector is a solution - but this fact need not bother us at present.) For each player i, one equation in (2.15) is redundant; it is automatically satisfied if the others are. If we select, for each player i, one strategy $\bar{a}_i \in A_i$ and delete the corresponding equation, we are left with $m = \sum_i |A_i| - I$ equations. Similarly we can delete the variable $s_i(\bar{a}_i)$ for each i as it can be recovered from the constraint that probabilities add up to one. Hence, (2.15) reduces to a system of m equations with m unknowns.

Taking each pair (i,a) with $i \in I$ and $a \in A_i \backslash \{\bar{a}_i\}$ as a coordinate, we can view S as a subset of R^m and the left-hand side of (2.15) as a mapping from S to R^m, hence

$$f_{ia_i}(s) = s_i(a_i)[u_i(s\backslash a_i) - u_i(s)] \qquad i \in I, a_i \in A_i \backslash \{\bar{a}_i\}. \tag{2.16}$$

Write $\partial f(s)$ for the Jacobian matrix of partial derivates of f evaluated at s and $|\partial f(s)|$ for its determinant. We say that s is a *regular equilibrium* of g if $|\partial f(s)| \neq 0$, hence, if the Jacobian is nonsingular. The reader easily checks that for all $i \in I$ and $a_i \in A_i$, if $s_i(a_i) = 0$, then $u_i(s\backslash a_i) - u_i(s)$ is an eigenvalue of $\partial f(s)$, hence, it follows that a regular equilibrium is necessarily quasi-strict. Furthermore, if s is a strict equilibrium, the above observation identifies m (hence, all) eigenvalues, so that any strict equilibrium is regular. A straightforward application of the implicit function theorem yields that, if s^* is a regular equilibrium of a game g^*, there exist neighborhoods U of g^* in Γ and V of s^* in S and a continuous map $s : U \to V$ with $s(g^*) = s^*$ and $\{s(g)\} = E(g) \cap V$ for all $g \in U$. Hence, if s^* is a regular equilibrium of g^*, then around (g^*, s^*) the equilibrium graph E looks like a continuous curve. By using Sard's theorem (in the manner initiated in Debreu (1970)) Harsanyi showed that for almost all normal form games all equilibria are regular. Formally, the proof proceeds by constructing a subspace $\tilde{\Gamma}$ of Γ and a polynomial map $\varphi : \tilde{\Gamma} \times S \to \Gamma$ with the following properties (where $\tilde{g}$ denotes the projection of g in $\tilde{\Gamma}$):

1. $\varphi(\tilde{g}, s) = g$ if $s \in E(g)$
2. $|\partial\varphi(\tilde{g}, s)| = 0$ if and only if $|\partial f(s)| = 0$.

Hence, if s is an irregular equilibrium of g, then g is a critical value of φ and Sard's theorem guarantees that the set of such critical values has measure zero. (For further details we refer to Harsanyi (1973a) and Van Damme (1987a).) We summarize the above discussion in the following Theorem.

Theorem 4 *(Harsanyi (1973a)). Almost all normal form games are regular, that is, they have only regular equilibria. Around a regular game, the equilibrium correspondence consists of a finite number of continuous functions. Any strict equilibrium is regular and any regular equilibrium is quasi-strict.*

Note that Theorem 4 may be of limited value for games given originally in extensive form. Any such nontrivial extensive form gives rise to a strategic form that is not in general position, hence, that is not regular. We will return to generic properties associated with extensive form games in Section 4. We will now show that the finiteness mentioned in Theorem 4 can be strengthened to oddness. Again we trace the footsteps of Harsanyi (1973a) with minor modifications as suggested by Ritzberger (1994), a paper that in turn builds on Dierker (1972).

Consider a regular game g and add to it a logarithmic penalty term so that the payoff to i resulting from s becomes

$$u_i^{\varepsilon}(s) = u_i(s) + \varepsilon \sum_{a_i \in A_i} ln s_i(a_i) \qquad (i \in I, s \in S). \tag{2.17}$$

Obviously, an equilibrium of this game has to be in completely mixed strategies. (Since the payoff function is not multilinear, (2.10) and (2.12) are no longer equivalent; by an equilibrium we mean a strategy vector satisfying (2.12) with u_i replaced by u_i^{ε}. It follows easily from Kakutani's theorem that an equilibrium exists.) Hence, the necessary and sufficient conditions for equilibrium are given by the first order conditions:

$$f_{ia_i}^{\varepsilon}(s) = f_{ia_i}(s) + \varepsilon(1 - |A_i| s_i(a_i)) = 0 \qquad i \in I, a_i \in A_i \backslash \{\bar{a}_i\}. \tag{2.18}$$

Because of the regularity of g, g has finitely many equilibria, say $s^1, ..., s^K$. The implicit function theorem tells us that for small ε, system (2.18) has at least K solutions $\{s^k(\varepsilon)\}_{k=1}^K$ with $s^k(\varepsilon) \to s^k$ as $\varepsilon \to 0$. In fact there must be exactly K solutions for small ε: Because of regularity there cannot be two solution curves converging to the same s^k, and if a solution curve remained bounded away from the set $\{s^1, ..., s^K\}$, then it would have a cluster point and this would be an equilibrium of g. However, the latter is impossible since we have assumed g to be regular. Hence, if ε is small, f^{ε} has exactly as many zero's as g has equilibria. An application of the Poincaré-Hopf Theorem for manifolds with boundary shows that each f^{ε} has an odd number of zero's, hence, g has an odd number of equilibria. (To apply the Poincaré-Hopf Theorem, take a smooth

approximation to the boundary of S, for example,

$$S(\delta) = \{s \in S; \prod_{a_i \in A_i} s_i(a_i) \geq \delta \text{ all } i\}. \tag{2.19}$$

Then the Euler characteristic of $S(\delta)$ is equal to 1 and, for fixed ε, if δ is sufficiently small, f^ε points outward at the boundary of $S(\delta)$.) To summarize, we have shown:

Theorem 5 *(Harsanyi (1973a), Wilson (1971), Rosenmüller (1971)). Generic strategic form games have an odd number of equilibria.*

Ritzberger notes that actually we can say a little more. Recall that the index of a zero s of f is defined as the sign of the determinant $|\partial f(s)|$. By the Poincaré-Hopf Theorem and the continuity of the determinant

$$\sum_{s \in E(g)} sgn|\partial f(s)| = 1. \tag{2.20}$$

It is easily seen that the index of a pure equilibrium is $+1$. Hence, if there are l pure equilibria, there must be at least $l - 1$ equilibria with index -1, and these must be mixed. This latter result was also established in Gul et al. (1993). In this paper, the authors construct a map g from the space of mixed strategies S into itself such that s is a fixed point of g if and only if s is a Nash equilibrium. They define an equilibrium s to be regular if it is quasi-strict and if $\det(I - g'(s)) \neq 0$. Using the result that the sum of the Lefschetz indices of the fixed points of a Lefschetz function is $+1$ and the observation that a pure equilibrium has index $+1$, they obtain their result that a regular game that has k pure equilibria must have at least $k - 1$ mixed ones. The authors also show that almost all games have only regular equilibria.

Recall that already Nash (1951) worked with a function f of which the fixed points correspond with the equilibria of the game. (See also the remark immediately below Theorem 1.) Nash's function is, however, different from that of Gul et al. (1993), and different from the function that we worked with in (2.15). This raises the question of whether the choice of the function matters. In recent work, Govindan and Wilson (2000) show that the anwer is no. These authors define a Nash map as a continuous function $f : \Gamma \times S \to S$ that has the property that for each fixed game g the induced

map $f_g : S \to S$ has as its fixed points the set of Nash equilibria of g. Given such a Nash map, the index $ind(C, f)$ of a component C of Nash equilibria of g is defined in the usual way (see Dold (1972)). The main result of Govindan and Wilson (2000) states that for any two Nash maps f, f' and any component C we have $ind(C, f) = ind(C, f')$. Furthermore, if the degree of a component, $\deg(C)$, is defined as the local degree of the projection map from the graph E of the equilibrium correspondence to the space of games (cf. Theorem 3), then $ind(C, f) = \deg(C)$ (see Govindan and Wilson (1997)).

2.4 Computation of equilibria: The 2-person case

The papers of Rosenmüller and Wilson mentioned in the previous theorem proved the generic oddness of the number of equilibria of a strategic form game in a completely different way than we did. These papers generalized the Lemke and Howson (1964) algorithm for the computation of equilibria in bimatrix games to n-person games. Lemke and Howson had already established the generic oddness of the number of equilibria for bimatrix games and the only difference between the 2-person case and the n-person case is that in the latter the pivotal steps involve nonlinear computations rather than the linear ones in the 2-person case. In this subsection we restrict ourselves to 2-person games and briefly outline the Lemke/Howson algorithm, thereby establishing another proof for Theorem 5 in the 2-person case. The discussion will be based upon Shapley (1974).

Let $g = \langle A, u \rangle$ be a 2-person game. The nondegeneracy condition that we will use to guarantee that the game is regular is

$$|C(s)| \;\geq\; |B(s)| \quad \text{for all } s \in S \tag{2.21}$$

This condition is clearly satisfied for almost all bimatrix games and indeed ensures that all equilibria are regular. We write $L(s_i)$ for the set of "labels" associated with $s_i \in S_i$

$$L(s_i) = A_i \backslash C_i(s_i) \cup B_j(s_i). \tag{2.22}$$

If $m_i = |A_i|$, then, by (2.21), the number of labels if s_i is at most m_i. We will be interested in the set N_i of those s_i that have exactly m_i labels. This set is finite: the

regularity condition (2.21) guarantees that for each set $L \subset A_1 \cup A_2$ with $|L| = m_i$ there is at most one $s_i \in S_i$ such that $L(s_i) = L$. Hence, the labelling identifies the strategy, so that the word label is appropriate. If $s_i \in N_i \backslash A_i$, then for each $a_i \in L(s_i)$ there exists (because of (2.21)) a unique ray in S_i emanating at s_i of points s_i' with $L(s_i') = L(s_i) \backslash \{a_i\}$, and moving in the direction of this ray we find a new point $s_i'' \in N_i$ after a finite distance. A similar remark applies to $s_i \in N_i \cap A_i$, except that in that case we cannot eliminate the label corresponding to $B_j(s_i)$. Consequently, we can construct a graph T_i with node set N_i that has m_i edges (of points s_i' with $|L(s_i')| = m_i - 1$) originating from each node in $N_i \backslash A_i$ and that has $m_i - 1$ edges originating from each node in $N_i \cap A_i$. We say that two nodes are adjacent if they are connected by an edge, hence, if they differ by one label.

Now consider the "product graph" T in the product set S: the set of nodes is $N = N_1 \times N_2$ and two nodes s, s' are adjacent if for some i $\ s_i = s_i'$ while for $j \neq i$ we have that s_j and s_j' are adjacent in N_j. For $s \in S$, write $L(s) = L(s_1) \cup L(s_2)$. Obviously, we have that $L(s) = A_1 \cup A_2$ if and only if s is a Nash equilibrium of g. Hence, equilibria correspond to fully labelled strategy vectors and the set of such vectors will be denoted by E. The regularity assumption (2.21) implies that $E \subset N$, hence, E is a finite set. For $a \in A_1 \cup A_2$ write N^a for the set of $s \in N$ that miss at most the label a. The observations made above imply the following fundamental lemma:

Lemma 1:

(i) If $\ s \in E, \ \ s_i = a$, then s is adjacent to no node in N^a

(ii) If $\ s \in E, \ \ s_i \neq a$, then s is adjacent to exactly one node in N^a

(iii) If $\ s \in N^a \backslash E, \ \ s_i = a$, then s is adjacent to exactly one node in N^a

(iv) If $\ s \in N^a \backslash E, \ \ s_i \neq a$, then s is adjacent to exactly two nodes in N^a

Proof:

(i) In this case s is a pure and strict equilibrium, hence, any move away from s eliminates labels other than a.

(ii) If s is a pure equilibrium, then the only move that eliminates only the label a is to increase the probability of a in T_i. If s_i is mixed, then (2.21) implies that s_j is mixed as well. We either have $s_i(a) = 0$ or $a \in B_i(s_j)$. In the first case the only move that eliminates only label a is one in T_i (increase the probability of a), in the second case it is the unique move in T_j away from the region where a is a best response.

(iii) The only possibility that this case allows is $s = (a, b)$ with b being the unique best response to a. Hence, if a' is the unique best response against b, the a' is the unique action that is labelled twice. The only possible move to an adjacent point in N^a is to increase the probability of a' in T_i.

(iv) Let b be the unique action that is labelled by both s_1 and s_2, hence $\{b\} = L(s_1) \cap L(s_2)$. Note that s_i is mixed. If s_j is mixed as well, then we can either drop b from $L(s_1)$ in T_i or drop b from $L(s_2)$ in T_j. This yields two different possibilities and these are the only ones. If s_j is pure, then $b \in A_i$ and the same argument applies. ¤

The lemma now implies that an equilibrium can be found by tracing a path of almost completely labelled strategy vectors in N^a, i.e. vectors that miss at most a. Start at the pure strategy pair (a, b) where b is the best response to a. If a is also the best response to b, we are done. If not, then we are in case (iii) of the lemma and we can follow a unique edge in N^a starting at (a, b). The next node s we encounter is one satisfying either condition (ii) of the lemma (and then we are done) or condition (iv). In the latter case, there are two edges of N^a at s. We came in via one route, hence there is only one way to continue. Proceeding in similar fashion, we encounter distinct nodes of type (iv) until we finally hit upon a node of type (ii). The latter must eventually happen since N^a has finitely many nodes.

The lemma also implies that the number of equilibria is odd. Consider an equilibrium s' different from the one found by the above construction. Condition (ii) from the lemma guarantees that this equilibrium is connected to exactly one node in N^a as in condition

(iv) of the lemma. We can now repeat the above constructive process until we end up at yet another equilibrium s''. Hence, all equilibria, except the distinguished one constructed above, appear in pairs: The total number of equilibria is odd.

Note that the algorithm described in this subsection offers no guarantee to find more than one equilibrium, let alone to find all equilibria. Shapley (1981) discusses a way of transforming the paths so as to get access to some of the previously inaccessible equilibria.

2.5 Purification of mixed strategy equilibria

In Section 2.1 we noted that mixed strategies can be interpreted both as acts of deliberate randomization as well as representations of players' beliefs. The former interpretation seems intuitively somewhat problematic; it may be hard to accept the idea of making an important decision on the basis of a toss of a coin. Mixed strategy equilibria also seem unstable: To optimize his payoff a player does not need to randomize; any pure strategy in the support is equally as good as the equilibrium strategy itself. The only reason a player randomizes is to keep the other players in equilibrium, but why would a player want to do this? Hence, equilibria in mixed strategies seem difficult to interpret (Aumann and Maschler (1972), Rubinstein (1991)).

Harsanyi (1973a) was the first to discuss the more convincing alternative interpretation of a mixed strategy of player i as a representation of the ignorance of the opponents as to what player i is actually going to do. Even though player i may follow a deterministic rule, the opponents may not be able to predict i's actions exactly, since i's decision might depend on information that the opponents can only assess probabilistically. Harsanyi argues that each player always has a tiny bit of private information about his own payoffs and he modifies the game accordingly. Such a slightly perturbed game admits equilibria in pure strategies and the (regular) mixed equilibria of the original unperturbed game may be interpreted as the limiting beliefs associated with these pure equilibria of the perturbed games. In this subsection we give Harsanyi's construction and state and illustrate his main result.

Let $g = \langle A, u \rangle$ be an I-person normal form game and, for each $i \in I$, let X_i be a random

vector taking values in R^A. Let $X = (X_i)_{i \in I}$ and assume that different components of X are stochastically independent. Let F_i be the distribution function of X_i and assume that F_i admits a continuously differentiable density f_i that is strictly positive on some ball Θ_i around zero in R^A (and 0 outside that ball). For $\varepsilon > 0$, write $g^\varepsilon(X)$ for the game described by the following rules:

(i) nature draws x from X

(ii) each player i is informed about his component x_i

(iii) simultaneously and independently each player i selects an action $a_i \in A_i$

(iv) each player i receives the payoff $u_i(a) + \varepsilon x_i(a)$, where a is the action combination resulting from (iii).

Note that, if ε is small, a player's payoff is close to the payoff from g with probability approximately 1. What a player will do in $g^\varepsilon(X)$ depends on his observation and on his beliefs about what the opponents will do. Note that these beliefs are independent of his observation and that, no matter what the beliefs might be, the player will be indifferent between two pure actions with probability zero. Hence, we may assume that each player i restricts himself to a pure strategy in $g^\varepsilon(X)$, i.e. to a map $\sigma_i : \Theta_i \to A_i$. (If a player is indifferent, he himself does not care what he does and his opponents do not care since they attach probability zero to this event.) Given a strategy vector σ^ε in $g^\varepsilon(X)$ and $a_i \in A_i$ write $\Theta_i^{a_i}(\sigma^\varepsilon)$ for the set of observations where σ_i^ε prescribes to play a_i. If a player $j \neq i$ believes i is playing σ_i^ε, then the probability that j assigns to i choosing a_i is

$$s_i^\varepsilon(a_i) = \int_{\Theta_i^{a_i}(\sigma^\varepsilon)} dF_i. \tag{2.23}$$

The mixed strategy vector $s^\varepsilon \in S$ determined by (2.23) will be called the vector of beliefs associated with the strategy vector σ^ε. Note that all opponents j of i have the same beliefs about player i since they base themselves on the same information. The strategy combination σ^ε is an equilibrium of $g^\varepsilon(X)$ if, for each player i, it assigns an optimal action at each observation, hence

$$\text{if } \; x_i \in \Theta_i^{a_i}(\sigma^\varepsilon), \text{then } \; a_i \in \arg\max[u_i(s^\varepsilon \backslash a_i) + \varepsilon x_i(s^\varepsilon \backslash a_i)]. \tag{2.24}$$

We can now state Harsanyi's theorem

Theorem 6 *(Harsanyi (1973b)). Let g be a regular normal form game and let the equilibria be $s^1, ..., s^K$. Then, for sufficiently small ε, the game $g^\varepsilon(X)$ has exactly K equilibrium belief vectors, say $s^1(\varepsilon)..., s^K(\varepsilon)$, and these are such that $\lim_{\varepsilon \to 0} s^k(\varepsilon) = s^k$ for all k. Furthermore, the equilibrium $\sigma^k(\varepsilon)$ underlying the belief vector $s^k(\varepsilon)$ can be taken to be pure.*

We will illustrate this theorem by means of a simple example, the game from Fig. 1. (The "t" stands for "tough", the "w" for "weak", the game is a variation of the battle of the sexes.) For analytical simplicity, we will perturb only one payoff for each player, as indicated in the diagram

	w_2	t_2
t_1	$1, u_2 + \varepsilon x_2$	$0, 0$
w_1	$u_1 + \varepsilon x_1, u_2 + \varepsilon x_2$	$u_1 + \varepsilon x_1, 1$

Figure 1: A perturbed game $g^\varepsilon(x_1, x_2)$ $(0 < u_1, u_2 < 1)$

The unperturbed game g ($\varepsilon = 0$ in Figure 1) has 3 equilibria, $(t_1, w_2), (w_1, t_2)$ and a mixed equilibrium in which each player i chooses t_i with probability $s_i = 1 - u_j$ $(i \neq j)$. The pure equilibria are strict, hence, it is easily seen that they can be approximated by equilibrium beliefs of the perturbed games in which the players have private information: If ε is small, then (t_i, w_j) is a strict equilibrium of $g^\varepsilon(x_1, x_2)$ for a set of (x_1, x_2)-values with large probability. Let us show how the mixed equilibrium of g can be approximated. If player i assigns probability s_j^ε to j playing t_j, then he prefers to play t_i if and only if

$$1 - s_j^\varepsilon > u_i + \varepsilon x_i. \tag{2.25}$$

Writing F_i for the distribution of X_i we have that the probability that j assigns to the event (2.25) is $F_i((s_j^\varepsilon - u_i)/\varepsilon)$, hence, to have an equilibrium of the perturbed game we

must have

$$s_i^\varepsilon = F_i((1 - s_j^\varepsilon - u_i)/\varepsilon) \qquad i, j \in \{1, 2\}, i \neq j. \tag{2.26}$$

Writing G_i for the inverse of F_i, we obtain the equivalent conditions

$$1 - s_j^\varepsilon - u_i - \varepsilon G_i(s_i^\varepsilon) = 0 \qquad i, j \in \{1, 2\}, i \neq j. \tag{2.27}$$

For $\varepsilon = 0$, the system of equations has the regular, completely mixed equilibrium of g as a solution, hence, the implicit function theorem implies that, for ε sufficiently small, there is exactly one solution $(s_1^\varepsilon, s_2^\varepsilon)$ of (2.27) with $s_i^\varepsilon \to 1 - u_j$ as $\varepsilon \to 0$. These beliefs are the ones mentioned in Theorem 6. A corresponding pure equilibrium strategy for each player i is: play w_i if $x_i \leq (1 - s_j^\varepsilon - u_i)/\varepsilon$ and play b_i otherwise.

For more results on purification of mixed strategy equilibria, we refer to Aumann et al. (1983), Milgrom and Weber (1985) and Radner and Rosenthal (1982). These papers consider the case where the private signals that players receive do not influence the payoffs and they address the question of how much randomness there should be in the environment in order to enable purification. In Section 5 we will show that completely different results are obtained if players make common noisy observations on the entire game: In this case even some strict equilibria cannot be approximated.

3 Backward induction equilibria in extensive form games

Selten (1965) pointed out that, in extensive form games, not every Nash equilibrium can be considered self-enforcing. Selten's basic example is similar to the game g from Figure 2, which has (l_1, l_2) and (r_1, r_2) as its two pure Nash equilibria. The equilibrium (l_1, l_2) is not self-enforcing. Since the game is noncooperative, player 2 has no ability to commit himself to l_2. If he is actually called upon to move, player 2 strictly prefers to play r_2, hence, being rational, he will indeed play r_2 in that case. Player 1 can foresee that player 2 will deviate to r_2 if he himself deviates to r_1, hence, it is in the interest of player 1 to deviate from an agreement on (l_1, l_2). Only an agreement on (r_1, r_2) is self-enforcing.

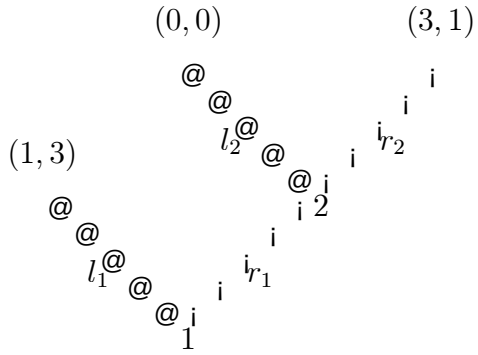


Figure 2: A Nash equilibrium that is not self-enforcing

Being a Nash equilibrium, (l_1, l_2) has the property that no player has an incentive to deviate from it if he expects the opponent to stick to this strategy pair. The example, however, shows that player 1's expectation that player 2 will abide by an agreement on (l_1, l_2) is nonsensical. For a self-enforcing agreement we should not only require that no player can profitably deviate if nobody else deviates, we should also require that the expectation that nobody deviates be rational. In this section we discuss several solution concepts, refinements of Nash equilibrium, that have been proposed as formalizations of this requirement. In particular, attention is focussed on sequential equilibria (Kreps and Wilson (1982a)) and on perfect equilibria (Selten (1975)). Along the way we will also discuss Myerson's (1978) notion of proper equilibrium. First, however, we introduce some basic concepts and notation related to extensive form games.

3.1 Extensive form and related normal forms

Throughout, attention will be confined to *finite extensive form games with perfect recall*. Such a game g is given by

(i) a collection I of players,

(ii) a game tree K specifying the physical order of play,

(iii) for each player i a collection H_i of information sets specifying the information a player has when he has to move. Hence H_i is a partition of the set of decision points of player i in the game and if two nodes x and y are in the same element h of the partition H_i, then i cannot distinguish between x and y,

(iv) for each information set h, a specification of the set of choices C_h that are feasible at that set,

(v) a specification of the probabilities associated with chance moves, and

(vi) for each end point z of the tree and each player i a payoff $u_i(z)$ that player i receives when z is reached.

For formal definitions, we refer to Selten (1975), Kreps and Wilson (1982a) or Hart (1992). For an extensive form game g we write $g = (\Gamma, u)$ where Γ specifies the structural characteristics of the game and u gives the payoffs. Γ is called a *game form*. The set of all games with game form Γ can be identified with an $|I| \times |Z|$ Euclidean space, where I is the player set and Z the set of end points. The assumption of *perfect recall*, saying that no player ever forgets what he has known or what he has done, implies that each H_i is a partially ordered set.

A *local strategy* s_{ih} of player i at $h \in H_i$ is a probability distribution on the set C_h of choices at this information set h. It is interpreted as a plan for what i will do at h or as the beliefs of the opponents of what i will do at that information set. Note that the latter interpretation assumes that different players hold the same beliefs about what i will do at h and that these beliefs do not change throughout the game. A *behavior strategy* s_i of player i assigns a local strategy s_{ih} to each $h \in H_i$. We write S_{ih} for the set of local strategies at h and S_i for the set of all behavior strategies of player i. A behavior strategy s_i is called pure if it associates a pure action at each $h \in H_i$ and the set of all these strategies is denoted A_i.

A behavior strategy combination $s = (s_1, ... \, s_I)$ specifies a behavior strategy for each player i. The probability distribution p^s that s induces on Z is called the *outcome* of s. Two strategies s_i' and s_i'' of player i are said to be *realization equivalent* if $p^{s \setminus s_i'} = p^{s \setminus s_i''}$ for each strategy combination s, i.e. if they induce the same outcomes against

any strategy profile of the opponents. Player i's *expected payoff* associated with s is $u_i(s) = \sum_z p^s(z)u_i(z)$. If x is a node of the game tree, then p^s_x denotes the probability distribution that results on Z when the game is started at x with strategies s and $u_{ix}(s)$ denotes the associated expectation of u_i. If every information set h of g that contains a node y after x actually has all its nodes after x, then that part of the tree of g that comes after x is a game of its own. It is called the *subgame* of g starting at x.

The *normal form* associated with g is the normal form game $\langle A, u\rangle$ which has the same player set, the same sets of pure strategies and the same payoff functions as g has. A mixed strategy from the normal form induces a behavioral strategy in the extensive form and Kuhn's (1953) theorem for games with perfect recall guarantees that, conversely, for every mixed strategy, there exists a behavior strategy that is realization equivalent to it. (See Hart (1992) for more details.) Note that the normal form frequently contains many realization equivalent pure strategies for each player: If the information set $h \in H_i$ is excluded by player i's own strategy, then it is "irrelevant" what the strategy prescribes at h. The game that results from the normal form if we replace each equivalence class (of realization equivalent) pure strategies by a representative from that class, will be called the *semi-reduced normal form*. Working with the semi-reduced normal form implies that we do not specify player$j$'s beliefs about what i will do at an information set $h \in H_i$ that is excluded by i's own strategy.

The *agent normal* form associated with g is the normal form game $\langle C, u\rangle$ that has a player ih associated with every information set h of each player i in g. This player ih has the set C_h of feasible actions as his pure strategy set and his payoff function is the payoff of the player i to whom he belongs. Hence, if $c_{ih} \in C_h$ for each $h \in \cup_i H_i$, then $s = (c_{ih})_{ih}$ is a (pure) strategy combination in g and we define $u_{ih}(s) = u_i(s)$ for $h \in H_i$. The agent normal form was first introduced in Selten (1975). It provides a local perspective, it decentralizes the strategy decision of player i into a number of local decisions. When planning his decision for h, the player does not necessarily assume that he is in full control of the decision at an information set $h' \in H_i$ that comes after h, but he is sure that the player/agent making the decision at that stage has the same objectives as he has. Hence, a player is replaced by a team of identically motivated

agents.

Note that a pure strategy combination is a Nash equilibrium of the agent normal form if and only if it is a Nash equilibrium of the normal form. Because of perfect recall, a similar remark applies to equilibria that involve randomization, provided that we identify strategies that are realization equivalent. Hence, we may define a Nash equilibrium of the extensive form as a Nash equilibrium of the associated (agent) normal form and obtain (2.12) as the defining equations for such an equilibrium. It follows from Theorem 1 that each extensive form game has at least one Nash equilibrium. Theorems 2 and 3 give information about the structure of the set of Nash equilibria of extensive form games. Kreps and Wilson proved a partial generalization of Theorem 4:

Theorem 7 *(Kreps and Wilson (1982a)). Let* Γ *be any game form. Then, for almost all* u*, the extensive form game* $\langle \Gamma, u \rangle$ *has finitely many Nash equilibrium outcomes (i.e. the set* $\{p^s(u) : s$ *is a Nash equilibrium of* $\langle \Gamma, u \rangle\}$ *is finite) and these outcomes depend continuously on* u*.*

Note that in this theorem, finiteness cannot be strengthened to oddness: Any extensive form game with the same structure as in Figure 2 and with payoffs close to those in Figure 2 has l_1 and (r_1, r_2) as Nash equilibrium outcomes. Hence, Theorem 5 does not hold for extensive form games. Little is known about whether Theorem 6 can be extended to classes of extensive form games. However, see Fudenberg et al. (1988) for results concerning various forms of payoff uncertainty in extensive form games.

Before moving on to discuss some refinements in the next subsections, we briefly mention some coarsenings of the Nash concept that have been proposed for extensive form games. Pearce (1984), Battigalli (1997) and Börgers (1991) propose concepts of extensive form rationalizability. Some of these also aim to capture some aspects of forward induction (see Section 4). Fudenberg and Levine (1993ab) and Rubinstein and Wolinsky (1994) introduce, respectively, the concepts of "self-confirming equilibria" and of "rationalizable conjectural equilibria" that impose restrictions that are in between those of Nash equilibrium and rationalizability. These concepts require players to hold identical and

correct beliefs about actions taken at information sets that are on the equilibrium path, but allow players to have different beliefs about opponents' play at information sets that are not reached. Hence, in such an equilibrium, if players only observe outcomes, no player will observe play that contradicts his predictions.

3.2 Subgame perfect equilibria

The Nash equilibrium condition (2.12) requires that each player's strategy be optimal from the ex ante point of view. Ex ante optimality implies that the strategy is also ex post optimal at each information set that is reached with positive probability in equilibrium, but, as the game of Figure 2 illustrates, such ex post optimality need not hold at the unreached information sets. The example suggests imposing ex post optimality as a necessary requirement for self-enforcingness but, of course, this requirement is meaningful only when conditional expected payoffs are well-defined, i.e. when the information set is a singleton. In particular, the suggestion is feasible for *games with perfect information*, i.e. games in which all information sets are singletons, and in this case one may require as a condition for s^* to be self-enforcing that it satisfies

$$u_{ih}(s^*) \geq u_{ih}(s^* \backslash s_i) \quad \text{for all } i, \text{ all } s_i \in S_i \text{ all } h \in H_i. \tag{3.1}$$

Condition (3.1) states that at no decision point h can a player gain by deviating from s^* if after h no other player deviates from s^*. Obviously, equilibria satisfying (3.1) can be found by rolling back the game tree in a dynamic programming fashion, a procedure already employed in Zermelo (1912). It is, however, also worthwhile to remark that already in Von Neumann and Morgenstern (1944) it was argued that this backward induction procedure was not necessarily justified as it incorporates a very strong assumption of "persistent" rationality. Recently, Hart (1999) has shown that the procedure may be justified in an evolutionary setting. Adopting Zermelo's procedure one sees that, for perfect information games, there exists at least one Nash equilibrium satisfying (3.1) and that, for generic perfect information games, (3.1) selects exactly one equilibrium. Furthermore, in the latter case, the outcome of this equilibrium is the unique outcome that survives iterated elimination of weakly dominated strategies in the normal form

of the game. (Each elimination order leaves at least this outcome and there exists a sequence of eliminations that leaves nothing but this outcome, cf. Moulin (1979).)

Selten (1978) was the first paper to show that the solution determined by (3.1) may be hard to accept as a guide to practical behavior. (Of course, it was already known for a long time that in some games, such as chess, playing as (3.1) dictates may be infeasible since the solution s^* cannot be computed.) Selten considered the finite repetition of the game from Figure 2, with one player 2 playing the game against a sequence of different players in each round and with players always being perfectly informed about the outcomes in previous rounds. In the story that Selten associates with this game, player 2 is the owner of a chain store who is threatened by entry in each of finitely many towns. When entry takes place (r_1 is chosen), the chain store owner either acquiesces (chooses r_2) or fights entry (chooses l_2). The backward induction solution has players play (r_1, r_2) in each round, but intuitively, we expect player 2 to behave aggressively (choose l_2) at the beginning of the game with the aim of inducing later entrants to stay out. The *chain store paradox* is the paradox that even people who accept the logical validity of the backward induction reasoning somehow remain unconvinced by it and do not act in the manner that it prescribes, but rather act according to the intuitive solution. Hence, there is an inconsistency between plausible human behavior and game-theoretic reasoning. Selten's conclusion from the paradox is that a theory of perfect rationality may be of limited relevance for actual human behavior and he proposes a theory of limited rationality to resolve the paradox. Other researchers have argued that the paradox may be caused more by the inadequacy of the model than by the solution concept that is applied to it. Our intuition for the chain store game may derive from a richer game in which the deterrence equilibrium indeed is a rational solution. Such richer models have been constructed in Kreps and Wilson (1982b), Milgrom and Roberts (1982) and Aumann (1992). These papers change the game by allowing a tiny probability that player 2 may actually find it optimal to fight entry, which has the consequence that, when the game still lasts for a long time, player 2 will always play as if it is optimal to fight entry which forces player 1 to stay out.

The cause of the chain store paradox is the assumption of *persistent rationality* that

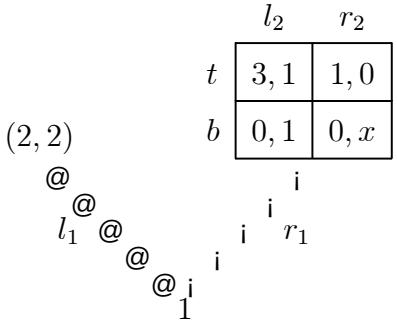
underlies (3.1), i.e. players are forced to believe that even at information sets h that can be reached only by many deviations from s^*, behavior will be in accordance with s^*. This assumption that forces a player to believe that an opponent is rational even after he has seen the opponent make irrational moves has been extensively discussed and criticized in the literature, with many contributions being critical (see, for example, Basu (1988, 1990), Ben Porath (1993), Binmore (1987), Reny (1992ab, 1993) and Rosenthal (1981)). Binmore argues that human rationality may differ in systematic ways from the perfect rationality that game theory assumes, and he urges theorists to build richer models that incorporate explicit human thinking processes and that take these systematic deviations into account. Reny argues that (3.1) assumes that there is common knowledge of rationality throughout the game, but that this assumption is self-contradicting: Once a player has "shown" that he is irrational (for example, by playing a strictly dominated move), rationality can no longer be common knowledge and solution concepts that build on this assumption are no longer appropriate. Aumann and Brandenburger (1995) however argue that Nash equilibrium does not build on this common knowledge assumption. Reny (1993), on the other hand, concludes from the above that a theory of rational behavior cannot be developed in a context that does not allow for irrational behavior, a conclusion similar to the one also reached in Selten (1975) and Aumann (1987b). Aumann (1995), however, disagrees with the view that the assumption of common knowledge of rationality is impossible to maintain in extensive form games with perfect information. As he writes, "The aim of this paper is to present a coherent formulation and proof of the principle that in PI games, common knowledge of rationality implies backward induction" (p. 7) (see also Aumann (1998) for an application to Rosenthal's centipede game; the references in that paper provide further information, also on other points of view).

We now leave this discussion on backward induction in games with perfect information and move on to discuss more general games. Selten (1965) notes that the argument leading to (3.1) can be extended beyond the class of games with perfect information. If the game g admits a subgame γ, then the expected payoffs of s^* in γ depend only on what s^* prescribes in γ. Denote this restriction of s^* to γ by s^*_γ. Once the subgame γ is reached,

all other parts of the game have become strategically irrelevant, hence, Selten argues that, for s^* to be self-enforcing, it is necessary that s^*_γ be self-enforcing for every subgame γ. Selten defined a *subgame perfect equilibrium* as an equilibrium s^* of g that induces a Nash equilibrium s^*_γ in each subgame γ of g and he proposed subgame perfection as a necessary requirement for self-enforcingness. Since every equilibrium of a subgame of a finite game can be "extended" to an equilibrium of the overall game, it follows that every finite extensive form game has at least one subgame perfect equilibrium.

Existence is, however, not as easily established for games in which the strategy spaces are continuous. In that case, not every subgame equilibrium is part of an overall equilibrium: Players moving later in the game may be forced to break ties in a certain way, in order to guarantee that players who moved earlier indeed played optimally. (As a simple example, let player 1 first choose $x \in [0,1]$ and let then player 2, knowing x, choose $y \in [0,1]$. Payoffs are give by $u_1(x,y) = xy$ and $u_2(x,y) = (1-x)y$. In the unique subgame perfect equilibrium both players choose 1 even though player 2 is completely indifferent when player 1 chooses $x = 1$.) Indeed, well-behaved continuous extensive form games need not have a subgame perfect equilibrium, as Harris et al. (1995) have shown. However, these authors also show that, for games with almost perfect information ("stage" games), existence can be restored if players can observe a common random signal before each new stage of the game which allows them to correlate their actions. For the special case where information is perfect, i.e. information sets are singletons, Harris (1985) shows that a subgame perfect equilibrium does exist even when correlation is not possible (see also Hellwig et al. (1990)).

Other chapters of this Handbook contain ample illustrations of the concept of subgame perfect equilibrium, hence, we will not give further examples. It suffices to remark here that subgame perfection is not sufficient for self-enforcingness, as is illustrated by the game from Figure 3.



	l_2	r_2
l_1	2, 2	2, 2
r_1t	3, 1	1, 0
r_1b	0, 1	0, x

Figure 3: Not all subgame perfect equilibria are self-enforcing

The left-hand side of Figure 3 illustrates a game where player 1 first chooses whether or not to play a 2×2 game. If player 1 chooses r_1, both players are informed that r_1 has been chosen and that they have to play the 2×2 game. This 2×2 game is a subgame of the overall game and it has (t, l_2) as its unique equilibrium. Consequently, (r_1t, l_2) is the unique subgame perfect equilibrium. The game on the right is the (semi-reduced) normal form of the game on the left. The only difference between the games is that, in the normal form, player 1 chooses simultaneously between l_1, r_1t and r_1b and that player 2 does not get to hear that player 1 has not chosen l_1. However, these changes appear inessential since player 2 is indifferent between l_2 and r_2 when player 1 chooses l_1. Hence, it would appear that an equilibrium is self-enforcing in one game only if it is self-enforcing in the other. However, the sets of subgame perfect equilibria of these games differ. The game on the right does not admit any proper subgames so that the Nash equilibrium (l_1, r_2) is trivially subgame perfect.

3.3 Perfect equilibria

We have seen that Nash equilibria may prescribe irrational, non-maximizing behavior at unreached information sets. Selten (1975) proposes to eliminate such non-self-enforcing equilibria by eliminating the possibility of unreached information sets. He proposes to look at complete rationality as a limiting case of incomplete rationality, i.e. to assume that players make mistakes with small vanishing probability and to restrict attention to

the limits of the corresponding equilibria. Such equilibria are called (trembling hand) perfect equilibria.

Formally, for an extensive form game g, Selten (1975) assumes that at each information set $h \in H_i$ player i will, with a small probability $\varepsilon_h > 0$, suffer from "momentary insanity" and make a mistake. Note that ε_h is assumed not to depend on the intended action at h. If such a mistake occurs, player i's behavior is assumed to be governed by some unspecified psychological mechanism which results in each choice c at h occurring with a strictly positive probability $\sigma_h(c)$. Selten assumes each of these probabilities ε_h and $\sigma_h(c) \quad (h \in H_i, c \in C_h)$ to be independent of each other and also to be independent of the corresponding probabilities of the other players. As a consequence of these assumptions, if a player i intends to play the behavior strategy s_i, he will actually play the behaviour strategy $s_i^{\varepsilon,\sigma}$ given by

$$s_i^{\varepsilon,\sigma}(c) = (1-\varepsilon_h)s_{ih}(c) + \varepsilon_h\sigma_h(c) \qquad (c \in C_h, h \in H_i). \tag{3.2}$$

Obviously, given these mistakes all information sets are reached with positive probability. Furthermore, if players intend to play $\bar{s}$, then, given the mistake technology specified by (ε, σ), each player i will at each information set h intend to choose a local strategy s_{ih} that satisfies

$$u_i(\bar{s}^{\varepsilon,\sigma} \backslash s_{ih}) \geq u_i(\bar{s}^{\varepsilon,\sigma} \backslash s'_{ih}) \qquad \text{all } \; s'_{ih} \in S_{ih}. \tag{3.3}$$

If (3.3) is satisfied by $s_{ih} = \bar{s}_{ih}$ at each $h \in \cup_i H_i$ (i.e. if the intended action optimizes the payoff taking the constraints into account), then $\bar{s}$ is said to be an equilibrium of the perturbed game $g^{\varepsilon,\sigma}$. Hence, (3.3) incorporates the assumption of persistent rationality. Players try to maximize whenever they have to move, but each time they fall short of the ideal. Note that the definitions have been chosen to guarantee that $\bar{s}$ is an equilibrium of $g^{\varepsilon,\sigma}$ if and only if $\bar{s}$ is an equilibrium of the corresponding perturbation of the agent normal form of g. A straightforward application of Kakutani's fixed point theorem yields that each perturbed game has at least one equilibrium. Selten (1975) then defines $\bar{s}$ to be a *perfect equilibrium* of g if there exist sequences ε^k, σ^k of mistake probabilities $(\varepsilon^k > 0, \varepsilon^k \to 0)$ and mistake vectors $\sigma^k_{ih}(c) > 0$ and an associated sequence s^k with s^k being an equilibrium of the perturbed game $g^{\varepsilon^k,\sigma^k}$ such that $s^k \to \bar{s}$ as $k \to \infty$. Since

the set of strategy vectors is compact, it follows that each game has at least one perfect equilibrium. It may also be verified that $\bar{s}$ is a perfect equilibrium of g if and only if there exists a sequence s^k of completely mixed behavior strategies ($s^k_{ih}(c) > 0$ for all i, h, c, k) that converges to $\bar{s}$ as $k \to \infty$, such that $\bar{s}_{ih}$ is a local best reply against any element in the sequence, i.e.

$$u_i(s^k \backslash \bar{s}_{ih}) = \max_{s_{ih} \in S_{ih}} u_i(s^k \backslash s_{ih}) \quad (\text{all } \ i, h, k). \tag{3.4}$$

Note that for $\bar{s}$ to be perfect, it is sufficient that $\bar{s}$ can be rationalized by some sequence of vanishing trembles, it is not necessary that $\bar{s}$ be robust against all possible trembles. In the next section we will discuss concepts that insist on such stronger stability. We will also encounter concepts that require robustness with respect to specific sequences of trembles. For example, Harsanyi and Selten's (1988) concept of uniformly perfect equilibria is based on the assumption that all mistakes are equally likely. In contrast, Myerson's (1978) properness concept builds on the assumption that mistakes that are more costly are much less likely.

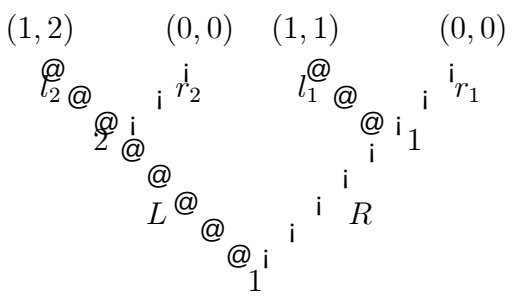
It is easily verified that each perfect equilibrium is subgame perfect. The converse is not true: In the game on the right of Figure 3 with $x \leq 1$, player 2 strictly prefers to play l_2 if player 1 chooses $r_1 t$ and $r_1 b$ by mistake, hence, only $(r_1 t, l_2)$ is perfect. However, since there are no subgames, (l_1, r_2) is subgame perfect.

By definition, the perfect equilibria of the extensive form game g are the perfect equilibria of the agent normal form of g. However, they need not coincide with the perfect equilibria of the associated normal form. Applying the above definitions to the normal form shows that $\bar{s}$ is a perfect equilibrium of a normal form game $g = \langle A, u \rangle$ if there exists a sequence of completely mixed strategy profiles s^k with $s^k \to \bar{s}$ such that $\bar{s} \in \mathcal{B}(s^k)$ for all k, i.e.

$$u_i(s^k \backslash \bar{s}_i) = \max_{s_i \in S_i} u_i(s^k \backslash s_i) \quad (\text{all } \ i, k). \tag{3.5}$$

Hence, we claim that the global conditions (3.5) may determine a different set of solutions than the local conditions (3.4). As a first example, consider the game from Figure 4. In

the extensive form, player 1 is justified to choose L if he expects himself, at his second decision node, to make mistakes with a larger probability than player 2 does. Hence, the outcome $(1, 2)$ is perfect in the extensive form. In the normal form, however, Rl_1 is a strategy that guarantees player 1 the payoff 1. This strategy dominates all others, so that perfectness forces player 1 to play it, hence, only the outcome $(1, 1)$ is perfect in the normal form. Motivated by the consideration that a player may be more concerned with mistakes of others than with his own, Van Damme (1984) introduces the concept of a quasi-perfect equilibrium. Here each player follows a strategy that at each node specifies an action that is optimal against mistakes of other players, keeping the player's own strategy fixed throughout the game. Mertens (1992) has argued that this concept of "quasi-perfect equilibria" is to be preferred above "extensive form perfect equilibria". (We will return to the concept below.)



	l_2	r_2
L	$1, 2$	$0, 0$
Rl_1	$1, 1$	$1, 1$
Rr_1	$0, 0$	$0, 0$

Figure 4: A perfect equilibrium of the extensive form need not be perfect in the normal form

Conversely, we have that a perfect equilibrium of the normal form need not even be subgame perfect in the extensive form. The game from Figure 3 with $x > 1$ provides an example. Only the outcome (3,1) is subgame perfect in the extensive form. In the normal form, player 2 is justified in playing r_2 if he expects that player 1 is (much) more likely to make the mistake $r_1 b$ than to make the mistake $r_1 t$. Hence, (l_1, r_2) is a perfect equilibrium in the normal form. Note that in both examples there is at least one equilibrium that is perfect in both the extensive and the normal form. Mertens (1992)

discusses an example in which the sets of perfect equilibria of these game forms are disjoint: the normal form game has a dominant strategy equilibrium, but this equilibrium is not perfect in the extensive form of the game.

It follows from (3.5) that a perfect equilibrium strategy of a normal form game cannot be weakly dominated. (Strategy s_i' is said to be *weakly dominated* by s_i'' if $u_i(s\backslash s_i'') \geq u_i(s\backslash s_i')$ for all s and $u_i(s\backslash s_i'') > u_i(s\backslash s_i')$ for some s.) Equilibria in undominated strategies are not necessarily perfect, but an application of the separating hyperplane theorem shows that the two concepts coincide in the 2-person case (Van Damme (1983)). (In the general case a strategy s_i is not weakly dominated if and only if it is a best reply against a completely mixed correlated strategy of the opponents.)

Before summarizing the discussion from this section in a theorem we note that games in which the strategy spaces are continua and payoffs are continuous need not have equilibria in undominated strategies. Consider the 2-player game in which each player i chooses x_i from $[0, \frac{1}{2}]$ and in which $u_i(x) = x_i$ if $x_i \leq x_j/2$ and $u_i(x) = x_j(1-x_i)/2 - x_j$ otherwise. Then the unique equilibrium is $x = 0$, but this is in dominated strategies. We refer to Simon and Stinchcombe (1995) for definitions of perfectness concepts for continuous games.

Theorem 8 *(Selten (1975)). Every game has at least one perfect equilibrium. Every extensive form perfect equilibrium is a subgame perfect equilibrium, hence, a Nash equilibrium. An equilibrium of an extensive form game is perfect if and only if it is perfect in the associated agent normal form. A perfect equilibrium of the normal form need not be perfect in the extensive form and also the converse need not be true. Every perfect equilibrium of a strategic form game is in undominated strategies and, in 2-person normal form games, every undominated equilibrium is perfect.*

3.4 Sequential equilibria

Kreps and Wilson (1982a) propose to eliminate irrational behavior at unreached information sets in a somewhat different way than Selten does. They propose to extend the applicability of (3.1) by explicitly specifying beliefs (i.e. conditional probabilities) at

each information set so that posterior expected payoffs can always be computed. Hence, whenever a player reaches an information set, he should, in conformity with Bayesian decision theory, be able to produce a probability distribution on the nodes in that set that represents his uncertainty. Of course, players' beliefs should be consistent with the strategies actually played (i.e. beliefs should be computed from Bayes' rule whenever possible) and they should respect the structure of the game (i.e. if a player has essentially the same information at h as at h', his beliefs at these sets should coincide). Kreps and Wilson ensure that these two conditions are satisfied by deriving the beliefs from a sequence of completely mixed strategies that converges to the strategy profile in question.

Formally, a *system of beliefs* μ is defined as a map that assigns to each information set $h \in \cup_i H_i$ a probability distribution μ_h on the nodes in that set. The interpretation is that, when $h \in H_i$ is reached, player i assigns a probability $\mu_h(x)$ to each node x in h. The system of beliefs μ is said to be *consistent* with the strategy profile s if there exists a sequence s^k of completely mixed behavior strategies ($s^k_{ih}(c) > 0$ for all i, h, k, c) with $s^k \to s$ as $k \to \infty$ such that

$$\mu_h(x) = \lim_{k \to \infty} p^{s^k}(x|h) \quad \text{for all } h, x \tag{3.6}$$

where $p^{s^k}(x|h)$ denotes the (well-defined) conditional probability that x is reached given that h is reached and s^k is played. Write $u^\mu_{ih}(s)$ for player i's expected payoff at h associated with s and μ, hence $u^\mu_{ih}(s) = \Sigma_{x \in h} \mu_h(x) u_{ix}(s)$, where u_{ix} is as defined in Section 3.1. The profile s is said to be *sequentially rational* given μ if

$$u^\mu_{ih}(s) \geq u^\mu_{ih}(s \backslash s'_i) \quad \text{ all } \ i, h, s'_i. \tag{3.7}$$

An assessment (s, μ) is said to be a *sequential equilibrium* if μ is consistent with s and if s is sequentially rational given μ. Hence, the difference between perfect equilibria and sequential equilibria is that the former concept requires ex post optimality approaching the limit, while the latter requires this only at the limit. Roughly speaking, perfectness amounts to sequentiality plus admissibility (i.e. the prescribed actions are not locally dominated). Hence, if s is perfect, then there exists some μ such that (s, μ) is a sequential equilibrium, but the converse does not hold: In a normal form game every Nash

equilibrium is sequential, but not every Nash equilibrium is perfect. The difference between the concepts is only marginal: for almost all games the concepts yield the same outcomes. The main innovation of the concept of sequential equilibrium is the explicit incorporation of the system of beliefs sustaining the strategies as part of the definition of equilibrium. In this, it provides a language for discussing the relative plausibility of various systems of beliefs and the associated equilibria sustained by them. This language has proved very effective in the discussion of equilibrium refinements in games with incomplete information (see, for example, Kreps and Sobel (1994)). We summarize the above remarks in the following theorem. (In it, we abuse the language somewhat: $s \in S$ is said to be a sequential equilibrium if there exists some μ such that (s, μ) is sequential.)

Theorem 9 *(Kreps and Wilson (1982a), Blume and Zame (1994)). Every perfect equilibrium is sequential and every sequential equilibrium is subgame perfect. For any game structure Γ we have that for almost all games $\langle \Gamma, u \rangle$ with that structure the sets of perfect and sequential equilibria coincide. For such generic payoffs u, the set of perfect equilibria depends continuously on u.*

Let us note that, if the action spaces are continua, and payoffs are continuous, a sequential equilibrium need not exist. A simple example is the following signalling game (Van Damme (1987b)). Nature first selects the type t of player 1, $t \in \{0, 2\}$ with both possibilities being equally likely. Next, player 1 chooses $x \in [0, 2]$ and thereafter player 2, knowing x but not knowing t, chooses $y \in [0, 2]$. Payoffs are $u_1(t, x, y) = (x - t)(y - t)$ and $u_2(t, x, y) = (1 - x)y$. If player 2 does not choose $y = 2 - t$ at $x = 1$, then type t of player 1 does not have a best response. Hence, there is at least one type that does not have a best response, and a sequential equilibrium does not exist.

In the literature one finds a variety of solution concepts that are related to the sequential equilibrium notion. In applications it might be difficult to construct an approximating sequence as in (3.6), hence, one may want to work with a more liberal concept that incorporates just the requirement that beliefs are consistent with s whenever possible,

hence $\mu_h(x) = p^s(s|h)$ whenever $p^s(h) > 0$. Combining this condition with the sequential rationality requirement (3.7) we obtain the concept of *perfect Bayesian equilibrium* which has frequently been applied in dynamic games with incomplete information. Some authors have argued that in the context of an incomplete information game, one should impose a support restriction on the beliefs: once a certain type of a player is assigned probability zero, the probability of this type should remain at zero for the remainder of the game. Obviously, this restriction comes in handy when doing backward induction. However, the restriction is not compelling and there may exist no Nash equilibria satisfying it (see Madrigal et al. (1987), Noldeke and Van Damme (1990)). For further discussions on variations of the concept of perfect Bayesian equilibrium, the reader is referred to Fudenberg and Tirole (1991).

Since the sequential rationality requirement (3.7) has already been discussed extensively in Section 3.2, there is no need to go into detail here. Rather let us focus on the consistency requirement (3.6). When motivating this requirement, Kreps/Wilson refer to the intuitive idea that when a player reaches an information set h with $p^s(h) = 0$, he reassesses the game, comes up with an alternative hypothesis s' (with $p^{s'}(h) > 0$) about how the game is played and then constructs his beliefs at h from s'. A system of beliefs is called structurally consistent if it can be constructed in this way. Kreps and Wilson claimed that consistency, as in (3.6), implies structural consistency, but this claim was shown to be incorrect in Kreps and Ramey (1987): There may not exist an equilibrium that can be sustained by beliefs that are both consistent and structurally consistent. At first sight this appears to be a serious blow to the concept of sequential equilibrium, or at least to its motivation. However, the problem may be seen to lie in the idea of reassessing the game, which is not intuitive at all. First of all, it goes counter to the idea of rational players who can foresee the play in advance: They would have to reassess at the start. Secondly, interpreting strategy vectors as beliefs about how the game will be played implies there is no reassessment: All agents have the same beliefs about the behavior of each agent. Thirdly, the combination of structural consistency with the sequential rationality requirement (3.7) is problematic: If player i believes at h that s' is played, shouldn't he then optimize against s' rather then against s? Of course, rejecting

structural consistency leaves us with the question of whether an alternative justification for (3.6) can be given. Kohlberg and Reny (1997) provide such a natural interpretation of consistency by relying on the idea of consistent probability systems.

3.5 Proper equilibria

In Section 3.1 we have seen that perfectness in the normal form is not sufficient to guarantee (subgame) perfectness in the extensive form. This observation raises the question of whether backward induction equilibria (say sequential equilibria) from the extensive form can already be detected in the normal form of the game. This question is important since it might be argued that, since a game is nothing but a collection of simultaneous individual decision problems, all information that is needed to solve these problems is already contained in the normal form of the game. The criteria for self-enforcingness in the normal form are no different from those in the extensive form: If the opponents of player i stick to s, then the essential information for i's decision problem is contained in this normal form: If i decides to deviate from s at a certain information set h, he can already plan that deviation beforehand, hence, he can deviate in the normal form. It turns out that the answer to the opening question is yes: An equilibrium that is proper in the normal form induces a sequential equilibrium outcome in every extensive form with that normal form.

Proper equilibria were introduced in Myerson (1978) with the aim of eliminating certain deficiencies in Selten's perfectness concept. One such deficiency is that adding strictly dominated strategies may enlarge the set of perfect equilibria. As an example, consider the game from the right-hand side of Figure 3 with the strategy r_1b eliminated. In this 2×2 game only (r_1t, b) is perfect. If we then add the strictly dominated strategy r_1b, the equilibrium (l_1, r_2) becomes perfect. But, of course, strictly dominated strategies should be irrelevant; they cannot determine whether or not an outcome is self-enforcing. Myerson argues that, in Figure 3, player 2 should not believe that the mistake r_1b is more likely than r_1t. On the contrary, since r_1t dominates r_1b, the mistake r_1b is more severe than the mistake r_1t; player 1 may be expected to spend more effort at preventing

it and as a consequence it will occur with smaller probability. In fact, Myerson's concept of proper equilibrium assumes such a more costly mistake to occur with a probability that is of smaller order.

Formally, for a normal form game $\langle A, u\rangle$ and some $\varepsilon > 0$, a strategy vector $s^\varepsilon \in S$ is said to be an ε-proper equilibrium if it is completely mixed (i.e. $s_i^\varepsilon(a_i) > 0$ for all i, all $a_i \in A_i$) and satisfies

$$\text{if } u_i(s^\varepsilon \backslash a_i) < u_i(s^\varepsilon \backslash b_i) \quad \text{then } s_i^\varepsilon(a_i) \leq \varepsilon s_i^\varepsilon(b_i) \quad (\text{all } i, a_i, b_i). \tag{3.8}$$

A strategy vector $s \in S$ is a *proper equilibrium* if it is a limit, as $\varepsilon \to 0$, of a sequence s^ε of $\varepsilon-$proper equilibria.

Myerson (1978) shows that each strategic form game has at least one proper equilibrium and it is easily seen that any such equilibrium is perfect. Now, let g be an extensive form game with semi-reduced normal form $n(g)$ and, for $\varepsilon \to 0$, let s^ε be an ε-proper equilibrium of $n(g)$ with $s^\varepsilon \to s$ as $\varepsilon \to 0$. Since s^ε is completely mixed, it induces a completely mixed behavior strategy $\bar{s}^\varepsilon$ in g. Let $\bar{s} = \lim_{\varepsilon \to 0} \bar{s}^\varepsilon$.Then $\bar{s}$ is a behavior strategy vector that induces the same outcome as s does, $p^{\bar{s}} = p^s$. (Note that s need not induce a full behavior strategy vector; as s was defined in the semi-reduced normal form, it does not necessarily specify a unique action at information sets that are excluded by the players themselves). Condition (3.8) now implies that at each information set h, $\bar{s}_i$ assigns positive probability only to the pure actions at h that maximize the local payoff at h against $\bar{s}^\varepsilon$. Namely, if c is a best response at h and c' is not, then for each pure strategy in the normal form that prescribes to play c' there exists a pure strategy that prescribes to play c and that performs strictly better against s^ε. (Take strategies that differ only at h.) Condition (3.8) then implies that in the normal form the total probability of the set of strategies choosing c' is of smaller order than the total probability of choosing c, hence, the limiting behavior strategy assigns probability 0 to c'. Hence, we have shown that each player always maximizes his local payoff, taking the mistakes of opponents into account. In other words, using the terminology of Van Damme (1984), the profile $\overline{s}$ is a quasi-perfect equilibrium. By the same argument, $\overline{s}$ is a sequential equilibrium. Formally, let μ^ε be the system of beliefs associated with $\bar{s}^\varepsilon$ and let $\mu = \lim_{\varepsilon \to 0} \mu^\varepsilon$. Then the assessment $(\bar{s}, \mu)$ satisfies (3.6) and (3.7), hence, it

is a sequential equilibrium of g. The following theorem summarizes the above discussion.

Theorem 10 *(i) (Myerson (1978)). Every strategic form game has at least one proper equilibrium. Every proper equilibrium is perfect.*

(ii) (Van Damme (1984), Kohlberg and Mertens (1986)). Let g be an extensive form game with semi-reduced normal form $n(g)$. If s is a proper equilibrium of $n(g)$, then p^s is a quasi-perfect and a sequential equilibrium outcome in g.

Mailath et al. (1997) have shown that sorts of converses to Theorem 10(ii) hold as well. Let $\{s^\varepsilon\}$ be a converging sequence of completely mixed strategies in a semi-reduced normal form game $n(g)$. This sequence induces a quasi-perfect equilibrium in every extensive form game with semi-reduced normal form $n(g)$ if and only if the limit of $\{s^\varepsilon\}$ is a proper equilibrium that is supported by the sequence. It is important that the same sequence be used: Hillas (1996) gives an example of a strategy profile that is not proper and yet is quasi-perfect in every associated extensive form. Secondly, Mailath et al. (1997) define a concept of normal form sequential equilibrium and they show that an equilibrium is normal form sequential if and only if it is sequential in every extensive form game with that semi-reduced normal form.

Theorem 10 (ii) appears to be the main application of proper equilibrium. One other application deserves to be mentioned: In 2-person zero-sum games, there is essentially one proper equilibrium and it is found by the procedure of cautious exploitation of the mistakes of the opponent that was proposed by Dresher (1961) (see Van Damme (1983, Sect. 3.5)).

4 Forward induction and stable sets of equilibria

Unfortunately, as the game of Figure 5 (a modification of a game discussed by Kohlberg (1981)) shows, none of the concepts discussed thus far provides sufficient conditions for self-enforcingness. In this game player 1 first chooses between taking up an outside

option that yields him 2 (and the opponent 0) and playing a battle-of-the-sexes game. Player 2 only has to move when player 1 chooses to play the subgame. In this game player 1 taking up his option and players continuing with (w_1, s_2) in the subgame constitutes a subgame perfect equilibrium. The equilibrium is even perfect: player 2 can argue that player 1 must have suffered from a sudden stroke of irrationality at his first move, but that his player will come back to his senses before his second move and continue with the plan (i.e. play w_1) as if nothing had happened. In fact, the equilibrium (t, s_2) is even proper in the normal form of the game: properness allows player 2 to conclude that the mistake pw_1 is more likely than the mistake ps_1 since pw_1 is better than ps_1 when player 2 plays s_2.

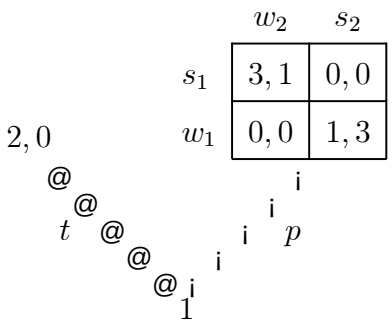


Figure 5: Battle of the sexes with an outside option

However, the outcome where player 1 takes up his option does not seem self-enforcing. If player 1 deviates and decides to play the battle-of-the-sexes game, player 2 should not rush to conclude that player 1 must have made a mistake; rather he might first investigate whether he can give a rational interpretation of this deviation. In the case at hand, such an explanation can indeed be given. For a rational player 1 it does not make sense to play w_1 in the subgame since the plan pw_1 is strictly dominated by the outside option. Hence, combining the rationality of player 1 with the fact that this player chose to play the subgame, player 2 should come to the conclusion that player 1 intends to play s_1 in the subgame, i.e. that player 1 bets on getting more than his option and that player 2 is sufficiently intelligent to understand this. Consequently, player 2

should respond by w_2, a move that makes the deviation of player 1 profitable, hence, the equilibrium is not self-enforcing.

Essentially what is involved here is an argument of forward induction: players' deductions about other players should be consistent with the assumption that these players are pursuing strategies that constitute rational plans for the overall game. The backward induction requirements discussed before were local requirements only taking into account rational behaviour in the future. Forward induction requires that players' deductions be based on overall rational behavior whenever possible and forces players to take a global perspective. Hence, one is led to an analysis by means of the normal form. In this section we take such a normal form perspective and ask how forward induction can be formulated. The discussion will be based on the seminal work of Elon Kohlberg and Jean-François Mertens (Kohlberg and Mertens (1986), Kohlberg (1989), Mertens (1987, 1989ab, 1991)). At this stage the reader may wonder whether there is no loss of information in moving to the normal form, i.e. whether the concepts that were discussed before can be recovered in the normal form. Theorem 10(ii) already provides part of the answer as it shows that sequential equilibria can be recovered. Mailath et al. (1993) discuss the question in detail and they show that also subgames and subgame perfect equilibria can be recovered in the normal form.

4.1 Set-valuedness as a consequence of desirable properties

Kohlberg and Mertens (1986) contains a first and partial axiomatic approach to the problem of what constitutes a self-enforcing agreement. (It should, however, be noted that the authors stress that their requirements should not be viewed as axioms since some of them are phrased in terms that are outside of decision theory.) Kohlberg and Mertens argue that a solution of a game should:

(i) always exist,

(ii) be consistent with standard one-person decision theory,

(iii) be independent of irrelevant alternatives, and

(iv) be consistent with backward induction.

(The third requirement states that strategies which certainly will not be used by rational players can have no influence on whether a solution is self-enforcing; it is the formalisation of the forward induction requirement that was informally discussed above; it will be given a more precise meaning below.) In this subsection we will discuss these requirements (except for (iv), which was extensively discussed in the previous section), and show that they imply that a solution cannot just be a single strategy profile but rather has to be a set of strategy profiles. In the next subsection, we give formalized versions of these basic requirements.

The existence requirement is fundamental and need not be discussed further. It guarantees that, if our necessary conditions for self-enforcingness leave only one candidate solution, that solution is indeed self-enforcing. Without having an existence theorem, we would run the risk of working with requirements that are incompatible, hence, of proving vacuous theorems.

The second requirement from the above list follows from the observation that a game is nothing but a simultaneous collection of one-person decision problems. In particular, it implies that the solution of a game can depend only on the normal form of that game. As a matter of fact, Kohlberg/Mertens argue that even less information than is contained in the normal form should be sufficient to decide on self-enforcingness. Namely, they take mixed strategies seriously as actions, and argue that a player is always able to add strategies that are just mixtures of strategies that are already explicitly given to them. Hence, they conclude that adding or deleting such strategies can have no influence on self-enforcingness. Formally, define the *reduced normal form* of a game as the game that results when all pure strategies that are equivalent to mixtures of other pure strategies have been deleted. (Hence, strategy $a_i \in A_i$ is deleted if there exists $s_i' \in S_i$ with $s_i'(a_i) = 0$ such that $u_j(s \backslash s_i') = u_j(s \backslash a_i)$ for all j. The reader may ask whether *the* reduced normal form is well-defined. We return to this issue in the next subsection.) As a first consequence of consistency with one-person decision theory, Kohlberg/Mertens insist that two games with the same reduced normal form be considered equivalent and, hence, as having the same solutions.

Kohlberg/Mertens accept as a basic postulate from standard decision theory that a rational agent will only choose undominated strategies, i.e. that he will not choose a strategy that is weakly dominated. Hence, a second consequence of (ii) is that game solutions should be undominated (admissible) as well. Furthermore, if players do not choose undominated strategies, such strategies are actually irrelevant alternatives, hence, (iii) requires that they can be deleted without changing the self-enforcingness of the solution. Hence, the combination of (ii) and (iii) implies that self-enforcing solutions should survive iterated elimination of weakly dominated strategies. Note that the requirement of independence of dominated strategies is a "global" requirement that is applicable independent of the specific game solution that is considered. Once one has a specific candidate solution, one can argue that, if the solution is self-enforcing, no player will use a strategy that is not a best response against the solution, and, hence, that such inferior strategies should be irrelevant for the study of the self-enforcingness of the solution. Consequently, Kohlberg/Mertens require as part of (iii) that a self-enforcing solution remains self-enforcing when a strategy that is not a best response to this solution is eliminated.

Note that "axioms" (ii) and (iii) force the conclusion that only (3,1) can be self-enforcing in the game of Figure 5: only this outcome survives iterated elimination of weakly dominated strategies. The same conclusion can also be obtained without using such iterative elimination: It follows from backward induction together with the requirement that the solution should depend only on the reduced normal form. Namely, add to the normal form of the game of Figure 5 the mixed strategy $m = \lambda t + (1-\lambda)s_1$ with $\frac{1}{2} < \lambda < 1$ as an explicit pure strategy. The resulting game can be viewed as the normal form associated with the extensive form game in which first player 1 decides between the outside option t and playing a subgame with strategy sets $\{s_1, w_1, m\}$ and $\{s_2, w_2\}$. This extensive form game is equivalent to the extensive form from Figure 5, hence, they should have the same solutions. However, the newly constructed game only has (3,1) as a subgame perfect equilibrium outcome. (In the subgame w_1 is strictly dominated by m, hence player 2 is forced to play w_2.)

We will now show why (subsets of) the above "axioms" can only be satisfied by set-valued solution concepts. Consider the trivial game from Figure 6. Obviously, player 1 choosing his outside option is self-enforcing. The question is whether a solution should contain a unique recommendation for player 2. Note that the unique subgame perfect equilibrium of the game requires player 2 to choose $\frac{1}{2}l_2 + \frac{1}{2}r_2$, hence, according to (i) and (iv) this strategy should be the solution for player 2. However, according to (i), (ii), and (iii), the solution should be l_2 (eliminate the strategies in the order pl_1, r_2), while according to these same axioms the solution should also be r_2 (take the elimination order pr_1, l_2). Hence, we see that, to guarantee existence, we have to allow for set-valued solution concepts. Furthermore, we see that, even with set-valued concepts, only weak versions of the axioms - that just require set inclusion - can be satisfied. Actually, these weak versions of the axioms imply that in this game all equilibria should belong to the solution. Namely, add to the normal form of the game of Figure 6 the mixed strategy $\lambda t + (1-\lambda)pl_1$ with $0 < \lambda < \frac{1}{2}$ as a pure strategy. Then the resulting game is the normal form of an extensive form game that has $\mu l_2 + (1-\mu)r_2$ with $\mu = (1-2\lambda)/(2-\lambda)$ as the unique subgame perfect equilibrium strategy for player 2. Hence, as λ moves through $(0, \frac{1}{2})$, we trace half of the equilibrium set of player 2, viz. the set $\{\mu l_2 + (1-\mu)r_2 : 0 < \mu < \frac{1}{2}\}$. The other half can be traced by adding the mixed strategy $\lambda t + (1-\lambda)pr_1$ in a similar way. Hence, axioms (i), (ii) and (iv) imply that all equilibrium strategies of player 2 belong to the solution.

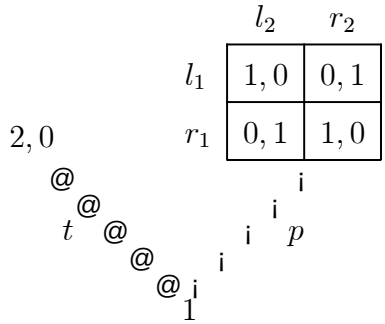
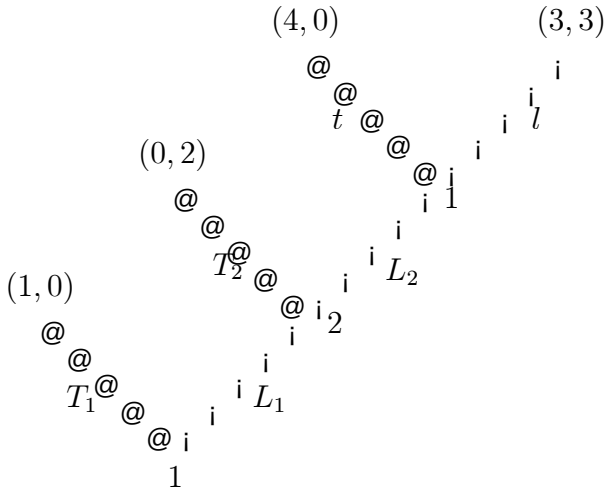


Figure 6: All Nash equilibria are equally good

The game of Figure 6 suggests that we should broaden our concept of a solution, that

we should not always aim to give a unique strategy recommendation for a player at an information set that can be reached only by irrational moves of other players. In the game of Figure 6, it is unnecessary to give a specific recommendation to player 2 and any such recommendation is somewhat artificial. Player 2 is dependent upon player 1 so that his optimal choice seems to depend on the exact way in which player 1 is irrational. However, our analysis has assumed rational players, and since no model of irrationality has been provided, the theorist could be content to remain silent. Hence, a self-enforcing solution should not necessarily pin down completely the behavior of players at unreached points of the game tree. We may be satisfied if we can recommend what players do in those circumstances that are consistent with players being rational, i.e. as long as the play is according to the self-enforcing solution.

Note that by extending our solution notion to allow for multiple beliefs and actions after irrational moves we can also get rid of the unattractive assumption of persistent rationality that was discussed in Section 3.2 and that corresponds to a narrow reading of axiom (iv). We might just insist that a solution contains a backward induction equilibrium, not that it consists exclusively of backward induction equilibria. We should not fully exclude the possibility that a player just made a one-time mistake and will continue to optimize, but we should not force this assumption. In fact, the axioms imply that the solution of a perfect information game frequently cannot just consist of the subgame perfect equilibrium. Namely, consider the game TOL(3) represented in Figure 7, which is a variation of a game discussed in Reny (1993). (TOL(n) stands for "Take it or leave it with n rounds". The game starts with \$1 on the table in round 1 and each time the game moves to a next round, the amount of money doubles. In round t, the player i with $i(\text{mod}2) = t(\text{mod}2)$ has to move. The game ends as soon as a player takes the money; if the game continues till the end, each player receives $\$2^{n-1} - 1$. (In the unique backwards induction equilibrium, player 1 takes the first dollar.)



	T_2	L_2
T_1	$1,0$	$1,0$
L_1t	$0,2$	$4,0$
L_1l	$0,2$	$3,3$

Figure 7: The game TOL (3)

The unique subgame perfect equilibrium of TOL(3) is (T_1t, T_2), which corresponds to (T_1, T_2) in the semi-reduced normal form. If the solution of the game were just (T_1, T_2), then L_1t would not be a best reply against the solution and according to "axiom" (iii), (T_1, T_2) should remain a solution when L_1t is eliminated. However, in the resulting 2×2 game, the unique perfect equilibrium is (L_1l, L_2) so that the axioms force this outcome to be the solution. Hence, the axioms imply that the strategy $\frac{1}{4}L_2 + \frac{3}{4}T_2$ of player 2 has to be part of the solution (in order to make L_1t a best response against the solution): player 1 cannot believe that, after player 2 has seen player 1 making the move L_1, player 2 believes player 1 to be rational. Intuitively, stable sets have to be large since they must incorporate the possibility of irrational play. Once we start eliminating dominated and/or inferior strategies, we attribute more rationality to players, make them more predictable and hence can make do with smaller stable sets. In formalizations of iterated elimination, we naturally have set inclusion.

The question remains of what type of mathematical objects are candidates for solutions of games now that we know that single strategy profiles do not qualify. In the above examples, the set of all equilibria was suggested, but the examples were special since there was only one connected component of Nash equilibria. More generally, one might consider connected components as solution candidates; however, this might be

too coarse. For example, if, in Figure 6, we were to change player 2's payoffs in the subgame in such a way as to make l_2 strictly dominant, we would certainly recommend player 2 to play l_2 even if player 1 has made the irrational move. Hence, the answer to the question appears unclear. Motivated by constructions like the above, and by the interpretation of stable sets as patterns in which equilibria vary smoothly with beliefs or presentation effects, Kohlberg/Mertens suggest connected subsets of equilibria as solution candidates. Hence, a solution is a subset of a component of the equilibrium set (cf. Theorem 2). Note that since a generic game has only finitely many Nash equilibrium outcomes (Theorem 7), all equilibria in the same connected component yield the same outcome (since outcomes depend continuously on strategies); hence, for generic games each Kohlberg/Mertens solution indeed generates a unique outcome. (See also Section 4.3.)

4.2 Desirable properties for strategic stability

In this subsection we rephrase and formalize (some consequences of) the requirements (i)-(iv) from the previous subsection, taking the discussion from that subsection into account.

Let Γ be the set of all finite games. A *solution concept* is a map $\mathcal{S}$ that assigns to each game $g \in \Gamma$ a collection of non-empty subsets of mixed strategy profiles for the game. A *solution* T of g is a subset T of the set of mixed strategies of (the normal form) of g with $T \in \mathcal{S}(g)$, hence, it is a set of profiles that $\mathcal{S}$ allows. The first fundamental requirement that we encountered in the previous subsection was:

(E) *Existence*: $\mathcal{S}(g) \neq \emptyset$

(We adopt the convention that, whenever a quantifier is missing, it should be read as "for all", hence (E) requires existence of at least one solution for each game.) Secondly, we will accept Nash equilibrium as a necessary requirement for self-enforcingness:

(NE) *Equilibrium*: If $T \in \mathcal{S}(g)$, then $T \subset E(g)$

A third requirement discussed above was

(C) *Connectedness*: If $T \in \mathcal{S}(g)$ then T is connected.

As discussed in the previous subsection, Kohlberg/Mertens insist that rational players only play admissible strategies. One formalization of admissibility is the restriction to undominated strategies, i.e. strategies that are best responses to correlated strategies of the opponents with full support. If players make their choices independently, a stronger admissibility suggests itself, viz. each player chooses a best response against a completely mixed strategy combination of the opponents. Formally, say that s_i' is an *admissible best reply* against s if there exists a sequence s^k of completely mixed strategy vectors converging to s such that s_i' is a best response against any element in the sequence. Write $\mathcal{B}_i^a(s)$ for the set of all such admissible best replies, $B_i^a(s) = \mathcal{B}_i^a(s) \cap A_i$, and let $\mathcal{B}^a(s) = X_i \mathcal{B}_i^a(s)$. For any subset S' of S write $\mathcal{B}^a(S') = \cup_{s \in S'} \mathcal{B}^a(s)$. We can now write the admissibility requirement as:

(A) *Admissibility*: If $T \in \mathcal{S}(g)$, then $T \subset \mathcal{B}^a(S)$.

Note that the combination of (NE) and (A) is almost equivalent to requiring perfection. The difference is that, as (3.5) shows, perfectness requires the approximating sequence s^k to be the same for each player. Accepting that players only use admissible best responses implies that a strategy that is not an admissible best response against the solution is certain not to be played and, hence, can be eliminated. Consequently, we can write the independence of irrelevant alternatives requirement as:

(IIA) *Independence of irrelevant alternatives*: If $a \notin B_i^a(T)$, then T contains a solution of the game in which a has been eliminated.

Note that $B_i^a(T) \subset B_i(T) \cap B_i^a(S)$, hence, (IIA) implies the requirements that strategies

that are not best responses against the solution can be eliminated and that strategies that are not admissible can be eliminated.

It is also a fundamental requirement that irrelevant players should have no influence on the solutions. Formally, following Mertens (1990), say that a subset J of the player set constitutes a *small world* if their payoffs do not depend on the actions of the players not in J, i.e.

$$\text{if } \; s_j = s'_j \; \text{ for all } \; j \in J, \; \text{ then } \; u_j(s) = u_j(s') \; \text{ for all } \; j \in J. \tag{4.1}$$

A solution has the small worlds property if the players outside the small world have no influence on the solutions inside the small world. Formally, if we write g_J for the game played by the insiders, then

(SMW) *Small worlds property*: If J is a small world in g, then T_J is a solution in g_J if and only if it is a projection of a solution T in g.

Closely related to the small worlds property is the decomposition property: If two disjoint player sets play different games in different rooms, it does not matter whether one analyses the games separately or jointly. Formally, say that g decomposes at J if both J and $\bar{J} = I \backslash J$ are small worlds in g.

(D) *Decomposition*: If g decomposes at J, then $T \in \mathcal{S}(g)$ if and only if $T = T_J \times T_{\bar{J}}$ with $T_k \in \mathcal{S}(g_k) \;\; (k \in \{J, \bar{J}\})$.

We now discuss the "player splitting property" which deals with another form of decomposition. Suppose g is an extensive form game and assume that there exists a partition P_i of H_i (the set of information sets of player i) such that, if h, h' belong to different elements of P_i, there is no path in the tree that cuts both h and h'. In such a case, the player can plan his actions at h without having to take into consideration his plans at h'. More generally, plans at one element of the partition can be made independently of plans at the other part and we do not limit the freedom of action of player i if we replace this player by a collection of agents, one agent for each element of P_i.

Consequently, we should require the two games to have the same self-enforcing solutions:

(PSP) *Player splitting property*: If g' is obtained from g by splitting some player i into a collection of independent agents, then $\mathcal{S}(g) = \mathcal{S}(g')$.

Note that for a solution concept having this property it does not matter whether a signalling game (Kreps and Sobel (1994)) is analysed in normal form (also called the Harsanyi-form in this case) or in agent normal form (also called the Selten-form). Also note that in (PSP) the restriction to independent agents is essential: In the agent normal form of the game from Figure 5, the first agent of player 1 taking up his outside option is a perfectly sensible outcome: Once the decisions are decoupled, the first action cannot signal anything about the second action. We will return to this in Section 5.
We will now formalize the requirement that the solution of a game depends only an those aspects of the problem that are relevant for the players' individual decision problems, i.e. that the solution is ordinal (cf. Mertens (1987)). As already discussed above, Mertens argues that rational players will only play admissible best responses. A natural invariance requirement thus is that the solutions depend only on the admissible best-reply correspondence, formally

(BRI) *Best reply invariance*: If $\mathcal{B}^a_g = \mathcal{B}^a_{g'}$ then, $\mathcal{S}(g) = \mathcal{S}(g')$.

Note that the application of (BRI) is restricted to games with the same player sets and the same strategy spaces, hence, this requirement should be supplemented with requirements that the names of the players and the strategies do not matter, etc.
In the previous subsection we also argued that games with the same reduced normal form should be considered equivalent. In order to be able to properly formalize this invariance requirement it turns out to be necessary to extend the domain of games somewhat: After one has eliminated all equivalent strategies of a player, this player's strategy set need no longer be a full simplex. To deal with such possibilities, define an I-person *strategic form game* as a tuple $\langle S, u\rangle$ where $S = X_i S_i$ is a product of compact

polyhedral sets and u is a multilinear map on S. Note that each such strategic form game has at least one equilibrium, and that the equilibrium set consists of finitely many connected components. Furthermore, all the requirements introduced above are meaningful for strategic form games. Say that an I-person strategic form game $g' = \langle S', u' \rangle$ is a *reduction* of the I-person normal form game $g = \langle A, u \rangle$ if there exists a map $f = (f_i)_{i \in I}$ with $f_i : S_i \to S_i'$ being linear and surjective, such that $u = u' \circ f$, hence, f preserves payoffs. Call such a map f an isomorphism from g onto g'. The requirement that the solution depends only on the reduced normal form may now be formalized as:

(I) *Invariance*: If f is an isomorphism from g onto g', then $\mathcal{S}(g') = \{f(T) : T \in \mathcal{S}(g)\}$ and $f^{-1}(T') = \cup\{T \in \mathcal{S}(g) : f(T) = T'\}$ for all $T' \in \mathcal{S}(g')$.

It should be stressed here that in Mertens (1987) the requirements (BRI) and (I) are derived from more abstract requirements of ordinality.
The final requirement that was discussed in the previous subsection was the backwards induction requirement, which, in view of Theorem 10, can be formalized as:

(BI) *Backwards induction*: If $T \in \mathcal{S}(g)$, then T contains a proper equilibrium of g.

4.3 Stable sets of equilibria

In Kohlberg and Mertens (1986), three set-valued solution concepts are introduced that aim to capture self-enforcingness. Unfortunately, each of these fails to satisfy at least one of the above requirements so that that seminal paper does not come up with a definite answer as to what constitutes a self-enforcing outcome. The definitions of these concepts build on Theorem 3 that describes the structure of the Nash equilibrium correspondence. The idea is to look at components of Nash equilibria that are robust to slight perturbations in the data of the game. The structure theorem implies that at least one such component exists. By varying the class of perturbations that are allowed, different concepts are obtained. Formally define
$(i) T$ is a *stable set of equilibria* of g if it is minimal among all the closed sets of equilibria

T' that have the property that each perturbed game $g^{\varepsilon,\sigma}$ with ε close to zero has an equilibrium close to T'.

$(ii)T$ is a *fully stable set of equilibria* of g if it is minimal among all the closed sets of equilibria T' that have the property that each game $\langle S', u\rangle$ with S'_i a polyhedral set in the interior of S'_i (for each i) that is close to g has an equilibrium close to T'.

$(iii)T$ is a *hyperstable set of equilibria* of g if it is minimal among all the closed sets of equilibria T' that have the property that for each game $g' = \langle A', u'\rangle$ that is equivalent to g and for each small payoff perturbation $\langle A', u^{\varepsilon}\rangle$ of g' there exists an equilibrium close to T'.

Kohlberg and Mertens (1986) show that every hyperstable set contains a set that is fully stable and that every fully stable set contains a stable set. Furthermore, from Theorem 3 they show that every game has a hyperstable set that is contained in a single connected component of Nash equilibria and, hence, that the same property holds for fully stable sets and stable sets. They, however, reject the (preliminary) concepts of hyperstability and full stability because these don't satisfy the admissibility requirement. Kohlberg/Mertens write that stability seems to be the "right" concept but they are forced to reject it since it violates (C) and (BI). (This concept does satisfy (E), (NE), (A), (IIA), (BRI), and (I).) Kohlberg/Mertens conclude with "we hope that in the future some appropriately modified definition of stability will, in addition, imply connectedness and backwards induction." Mertens (1989a, 1991) gives such a modification. We will consider it below.

An example of a game in which every fully stable set contains an inadmissible equilibrium (and hence in which every hyperstable set contains such an equilibrium) is obtained by changing the payoff vector $(0, 2)$ in TOL(3) (Figure 7) to $(5, 5)$. The unique admissible equilibrium then is (L_1t, T_2) but every fully stable set has to contain the strategy (L_1l) of player 1. Namely, if (in the normal form) player 1 trembles with a larger probability to T_1 when playing L_1t than when playing L_1l, we obtain a perturbed game in which only (L_1l, T_2) is an equilibrium.

We now describe a 3-person game (attributed to Faruk Gul in Kohlberg and Mertens (1986)) that shows that stable sets may contain elements from different equilibrium

components and need not contain a subgame perfect equilibrium. Player 3 starts the game by choosing between an outside option T (which yields payoffs $(0,0,2)$) or playing a simultaneous move subgame with players 1 and 2 in which each of the three players has strategy set $\{a,b\}$ and in which the payoffs are as in the matrix from the left-hand side of Figure 8 ($x,y \in \{a,b\}, x \neq y$). Hence, players 1 and 2 have identical payoffs and they want to make the same choice as player 3. Player 3 prefers these players to make different choices, but, if they make the same choice, he wants his choice to be different from theirs.

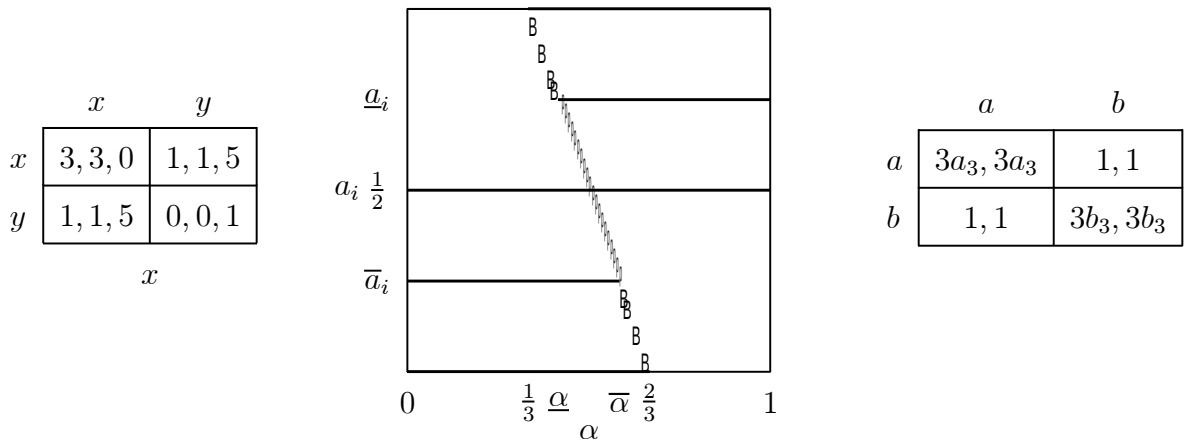


Figure 8: Stable sets need not contain a subgame perfect equilibrium

The game g described in the above story has a unique subgame perfect equilibrium: player 3 chooses to play the subgame and each player chooses $\frac{1}{2}a + \frac{1}{2}b$ in this subgame. This strategy vector constitutes a singleton component of the set of Nash equilibria. In addition, there are two components in which player 3 takes up his option T. Writing a_i (resp. b_i) for the probability with which player i ($i = 1, 2$) chooses a (resp. b), the strategies of players 1 and 2 in this component are the solutions to the pair of inequalities

$$4(a_1 + a_2) - 9a_1a_2 \leq 1 \text{ and } 4(b_1 + b_2) - 9b_1b_2 \leq 1. \tag{4.2}$$

Note that the solution set of (4.2) indeed consists of two connected components, one around (a,a) (i.e. $a_1 = a_2 = 1$) and one around (b,b). Now, let us look at perturbations

of (the normal form of) g. If player 3 chooses to play the subgame with positive stability ε and if, conditional on such a mistake, he chooses a (resp. b) with probability a_3 (resp. b_3), players 1 and 2 face the game from the right-hand side of Figure 8. The equilibria of this game are given by

$$a_i = \begin{cases} 0 & \text{if } \ a_3 < \frac{1}{3} \\ 0, 2 - 3a_3, 1 & \text{if } \ \frac{1}{3} < a_3 < \frac{2}{3} \\ 1 & \text{if } \ \frac{2}{3} < a_3, \end{cases}$$

hence, restricted to players 1 and 2, each perturbed game has (a, a) or (b, b) (or both) as a strict equilibrium. If players 1 and 2 coordinate on any of these strict equilibria, player 3 strictly prefers to play T, hence, $\{(a_1, a_2, T), (b_1, b_2, T)\}$ is a stable set of g. Obviously, this set does not contain the subgame perfect equilibrium, and even yields a different outcome.

A closer investigation may reveal the source of the difficulty and suggest a resolution of the problem. Since problematic zero-probability events arise only from player 3 choosing T, let us insist that he chooses to play the subgame with probability ε but, for simplicity, let us not perturb the strategies of players 1 and 2. Formally, consider a perturbed game $g^{\varepsilon,\sigma}$ with $\varepsilon_1 = \varepsilon_2 = 0, \varepsilon_3 = \varepsilon > 0$ and $\sigma_3 = (0, \alpha, 1 - \alpha)$, hence α is the probability that player 3 chooses a if he makes a mistake. The middle panel in Figure 8 displays, for any small $\varepsilon > 0$, the equilibrium correspondence as a function of α. (The horizontal axis corresponds to α, the vertical one to a_i.) Each perturbed game has an equilibrium close to the subgame perfect equilibrium of g. This equilibrium is represented by the horizontal line at $a_i = \frac{1}{2}$. The inverted z-shaped figure corresponds to the solutions of (4.3). If players 1 and 2 play such a solution that is sufficiently close to a pure strategy, then T is the unique best response of player 3, hence, in that case we have an equilibrium of the perturbed game with $a_3 = \alpha$. If players 1 and 2 play a solution of (4.3) that is sufficiently close to $a_i = \frac{1}{2}$ (i.e. they choose $a_i \in (\bar{a}_i, \underline{a}_i)$ corresponding to the dashed part of the z-curve), then we do not have an equilibrium unless $a_1 = a_2 = a_3 = \frac{1}{2}$. (If $a_i > \frac{1}{2}$, then the unique best response of player 3 is to play b, hence $a_3 = \varepsilon\alpha < \frac{1}{3}$ so that by (4.3) we should have $a_i = 0$.) The points $\underline{\alpha}$ and $\bar{\alpha}$ where the solid z-curve changes

into the dashed z-curve are somewhat special. Writing $\underline{a}_i = 2 - 3\underline{\alpha}$ we have that if each player i $(i = 1, 2)$ chooses a with probability $\underline{a}_i$, then player 3's best responses are T and b. Consequently, by playing b voluntarily with the appropriate probability, player 3 can enforce any $a_3 \in (\varepsilon\alpha, \alpha)$, hence, if ε is sufficiently small and $\alpha > \underline{\alpha}$, player 3 can enforce $a_3 = \underline{\alpha}$. We see that for each $\alpha \geq \underline{\alpha}$, the perturbed game has an equilibrium with $a_i = \underline{a}_i$. In the diagram, this branch is represented by the horizontal line at $\underline{a}_i$. Of course, there is a similar branch at $\bar{a}_i$. Since the above search was exhaustive, the middle panel in Figure 8 contains a complete description of the equilibrium graph, or at least of its projection on the (α, a_i)-space.

The critical difference between the "middle" branch of the equilibrium correspondence and each of the other two branches is that in the latter cases it is possible to continuously deform the graph, leaving the part over the extreme perturbations ($\alpha \in \{0, 1\}$) intact, in such a way that the interior is no longer covered, i.e. such that there are no longer "equilibria" above the positive perturbations. Hence, although the projection from the union of the top and bottom branches to the perturbations is surjective (as required by stability), this projection is homologically trivial, i.e. it is homologous to the identity map of the boundary of the space of perturbations. Building on this observation, and on the topological structure of the equilibrium correspondence more generally, Mertens (1989a, 1991) proposes a refinement of stability (to be called M-stability) that essentially requires that the projection from a neighborhood of the set to a neighborhood of the game should be homologically nontrivial. As the formal definition is somewhat involved we will not give it here but confine ourselves to stating its main properties. Let us, however, note that Mertens does not insist on minimality; he shows that this conflicts with the ordinality requirement (cf. Section 4.5).

Theorem 11 *(Mertens (1989a, 1990, 1991)). M-stable sets are closed sets of normal form perfect equilibria that satisfy all properties listed in the previous subsection.*

We close this subsection with a remark and with some references to recent literature. First of all, we note that also in Hillas (1990) a concept of stability is defined that satisfies

all properties from the list of the previous subsection (We will refer to this concept as H-stability. To avoid confusion, we will refer to the stability concept that was defined in Kohlberg and Mertens as KM-stability.) T is an *H-stable set of equilibria* of g if it is minimal among all the closed sets of equilibria T' that have the following property: each upper-hemicontinuous compact convex-valued correspondence that is pointwise close to the best-reply correspondence of a game that is equivalent to g has a fixed point close to T'. The solution concept of H-stable sets satisfies the requirements (E), (NE), (C), (A), (IIA), (BRI), (I) and (BI), but it does not satisfy the other requirement from Section 4.2. (The minimality requirement forces H-stable sets to be connected, hence, in the game of Figure 8 only the subgame perfect equilibrium outcome is H-stable.) In Hillas et al. (1999) it is shown that each M-stable set contains an H-stable set. That paper discusses a couple of other related concepts as well.

I conclude this Section by referring to some other recent work. Wilson (1997) discusses the role of admissibility in identifying self-enforcing outcomes. He argues that admissibility criteria should be deleted when selecting among equilibrium components, but that they may be used in selecting equilibria from a component, hence, Wilson argues in favour of perfect equilibria in essential components, i.e. components for which the degree (cf. Section 2.3) is non-zero. Govindan and Wilson (1999) show that, in 2-player games, maximal M-stable sets are connected components of perfect equilibria, hence, such sets are relatively easy to compute and their number is finite (On finiteness, see Hillas et al. (1997).) The result implies that an essential component contains a stable set, however, as Govindan/Wilson illustrate by means of several examples, inessential components may contain stable sets as well.

4.4 Applications of stability criteria

Concepts related to strategic stability have been frequently used to narrow down the number of equilibrium outcomes in games arising in economic contexts. (Recall that in generic extensive games all equilibria in the same component have the same outcome so that we can speak of stable and unstable outcomes.) Especially in the context of signalling games many refinements have been proposed that were inspired by stability

or by its properties (cf. Cho and Kreps (1987), Banks and Sobel (1987) and Cho and Sobel (1990)). As this literature is surveyed in the chapter by Kreps and Sobel (1994), there is no need to discuss these applications here (see Van Damme (1992)). I'll confine myself here to some easy applications and to some remarks on examples where the fine details of the definitions make the difference.

It is frequently argued that the Folk Theorem, i.e. the fact that repeated games have a plethora of equilibrium outcomes (see chapter 4 in this Handbook) shows a fundamental weakness of game theory. However, in a repeated game only few outcomes may actually be strategically stable. (General results, however, are not yet available.) To illustrate, consider the twice-repeated battle-of-the-sexes game, where the stage game payoffs are as in (the subgame occurring in) Figure 5 and that is played according to the standard information conditions. The path $\langle (s_1, w_2), (s_1, w_2) \rangle$ in which player 1's most preferred stage equilibrium is played twice is not stable. Namely, the strategy $s_2 w_2$ (i.e. deviate to s_2 and then play w_2) is not a best response against any equilibrium that supports this path, hence, if the path were stable, then according to (IIA) it should be possible to delete this strategy. However, the resulting game does not have an admissible equilibrium with payoff $(6, 2)$ so that the path cannot be stable. (Admissibility forces player 1 to respond with w_1 after 2 has played s_2; hence, the deviation $s_2 s_2$ is profitable for player 2.) For further results on stability in repeated games, the reader is referred to Balkenborg (1993), Osborne (1990), Ponssard (1991) and Van Damme (1989a).

Stability implies that the possibility to inflict damage on oneself confers power. Suppose that before playing the one-shot battle-of-the-sexes game, player 1 has the opportunity to burn 1 unit of utility in a way that is observable to player 2. Then the only stable outcome is the one in which player 1 does not burn utility and players play (s_1, w_1), hence, player 1 gets his most preferred outcome. The argument is simply that the game can be reduced to this outcome by using (IIA). If both players can throw away utility, then stability forces utility to be thrown away with positive probability: Any other outcome can be upset by (IIA). (See Van Damme (1989a) for further details and Ben Porath and Dekel (1992), Bagwell and Ramey (1996), Glazer and Weiss (1990) for applications.)

Most applications of stability in economics use the requirements from Section 4.2 to

limit the set of solution candidates to one and they then rely on the existence theorem to conclude that the remaining solution must be stable. Direct verification of stability may be difficult; one may have to enumerate all perturbed games and investigate how the equilibrium graph hangs together (see Mertens (1987, 1989a,b, 1991) for various illustrations of this procedure and for arguments as to why certain shortcuts may not work). Recently, Wilson (1992) has constructed an algorithm to compute a simply stable component of equilibria in bimatrix games. Simply stable sets are robust against a restricted set of perturbations, viz. one perturbs only one strategy (either its probability or its payoff). Wilson amends the Lemke/Howson algorithm from Section 2.3 to make it applicable to nongeneric bimatrices and he adds a second stage to it to ensure that it can only terminate at a simply stable set. Whenever the Lemke/Howson algorithm terminates with an equilibrium that is not strict, Wilson uses a perturbation to transit onto another path. The algorithm terminates only when all perturbations have been covered by some vertex in the same component. Unfortunately, Wilson cannot guarantee that a simply stable component is actually stable.

In Van Damme (1989a) it was argued that stable sets (as originally defined by Kohlberg/ Mertens) may not fully capture the logic of forward induction. Following an idea originally discussed in McLennan (1985) it was argued that if an information set $h \in H_i$ can be reached only by one equilibrium s^*, and if s^* is self-enforcing, player i should indeed believe that s^* is played if h is reached and, hence, only s^*_{ih} should be allowed at h. A 2-person example in Van Damme (1989a) showed that stable equilibria need not satisfy this forward-induction requirement. (Actually Gul's example (Figure 8) already shows this.) Hauk and Hurkens (1999) have recently shown that this forward-induction property is satisfied by none of the stability concepts discussed above. On the other hand they show that this property is satisfied by some evolutionary equilibrium concepts that are related to those discussed in Section 4.5 below.

Gul and Pearce (1996) argue that forward induction loses much of its power when public randomization is allowed; however, Govindan and Robson (1998) show that the Gul/Pearce argument depends essentially on the use of inadmissible strategies.

Mertens (1992) describes a game in which each player has a unique dominant strategy,

yet the pair of these dominant strategies is not perfect in the agent normal form. Hence, the M-stable sets of the normal form and those of the agent normal form may be disjoint. That same paper also contains an example of a nongeneric perfect information game (where ties are not noticed when doing the backwards induction where the unique M-stable set contains other outcomes besides the backwards induction outcome). (See also Van Damme (1987b), pp. 32-33.)

Govindan (1995) has applied the concept of M-stability to the Kreps and Wilson (1982b) chain store game with incomplete information. He shows that only the outcome that was already identified in Kreps and Wilson (1982b) as the unique "reasonable" one, is indeed the unique M-stable outcome. Govindan's approach is to be preferred to Kreps and Wilson's since it does not rely on ad hoc methods. It is worth remarking that Govindan is able to reach his conclusion just by using the properties of M-stable equilibria (as mentioned in Theorem 11) and that the connectedness requirement plays an important role in the proof.

4.5 Robustness and persistent equilibria

Many game theorists are not convinced that equilibria in mixed strategies should be treated on equal footing with pure, strict equilibria; they express a clear preference for pure equilibria. For example, Harsanyi and Selten (1988, p. 198) write, "Games that arise in the context of economic theory often have many strict equilibrium points. Obviously in such cases it is more natural to select a strict equilibrium point rather than a weak one. Of course, strict equilibrium points are not always available (...) but it is still possible to look for a principle that helps us to avoid those weak equilibrium points that are especially unstable." (They use the term "strong" where I write "strict"). In this subsection we discuss such principles.

Harsanyi and Selten discuss two forms of instability associated with mixed strategy equilibria. The first, weak form of instability results from the fact that even though a player might have no incentive to deviate from a mixed equilibrium, he has no positive incentive to play the equilibrium strategy either: any pure strategy that is used with positive probability is equally good. As we have seen in Section 2.5, the reinterpretation

of mixed equilibria as equilibria in beliefs provides an adequate response to the criticism that is based on this form of instability. The second, strong form of instability is more serious and cannot be countered so easily. This form of instability results from the fact that, in a mixed equilibrium, if a player's beliefs differ even slightly from the equilibrium beliefs, optimizing behavior will typically force the player to deviate from the mixed equilibrium strategy. In contrast, if an equilibrium is strict, a player is forced to play his equilibrium strategy as long as he assigns a sufficiently high probability to the opponents playing this equilibrium. For example, in the battle-of-the-sexes game (that occurs as the subgame in Figure 5), each player is willing to follow the recommendation to play a pure equilibrium as long as he believes that the opponent follows the recommendation with a probability of at least $\frac{2}{3}$. In contrast, player i is indifferent between s_i and w_i only if he assigns a probability of exactly $\frac{1}{3}$ to the opponent playing w_j. Hence, it seems that strict equilibria possess a type of robustness property that the mixed equilibrium lacks. However, this difference is not picked up by any of the stability concepts that have been discussed above: The mixed strategy equilibrium of the battle-of-the-sexes game constitutes a singleton stable set according to each of the above stability definitions. In this subsection, we will discuss some set-valued generalizations of strict equilibria that do pick up the difference. They all aim at capturing the idea that equilibria should be robust to small trembles in the equilibrium beliefs, hence, they address the question of what outcome an outsider would predict who is quite sure, but not completely sure, about the players' beliefs. The discussion that follows is inspired by Balkenborg (1992). If s is a strict equilibrium of $g = \langle A, u \rangle$, then s is the unique best response against s, hence $\{s\} = B(s)$. We have already encountered a set-valued analogue of this uniqueness requirement in Section 2.2, viz. the concept of a minimal curb set. Recall that $C \subset A$ is a curb set of g if

$$B(C) \subset C, \tag{4.3}$$

i.e. if every best reply against beliefs that are concentrated on C again belongs to C. Obviously, a singleton set C satisfies (4.4) only if it is a strict equilibrium. Nonsingleton curb sets may be very large (for example, the set A of all strategy profiles trivially satisfies (4.4)), hence in order to obtain more definite predictions, one can investigate

minimal sets with the property (4.4). In Section 2.2 we showed that such minimal curb sets exist, that they are tight, i.e. $B(C) = C$, and that distinct minimal curb sets are disjoint. Furthermore, curb sets possess the same neighborhood stability property as strict equilibria, viz. if C satisfies (4.4), then there exists a neighborhood U of $X_i\Delta(C_i)$ in S such that

$$B(U) \subset C. \tag{4.4}$$

Despite all these nice properties, minimal curb sets do not seem to be the appropriate generalization of strict equilibria. First, if a player i has payoff equivalent strategies, then (4.4) requires all of these to be present as soon as one is present in the set, but optimizing behavior certainly doesn't force this conclusion: It is sufficient to have at least one member of the equivalence class in the curb set. (Formally, define the strategies s_i' and s_i'' of player i to be *i-equivalent* if $u_i(s\backslash s_i') = u_i(s\backslash s_i'')$ for all $s \in S$, and write $s_i' \sim_i s_i''$ if s_i' and s_i'' are i-equivalent.) Secondly, requirement (4.4) does not differentiate among best responses, it might be preferable to work with the narrower set of admissible best responses. As a consequence of these two observations, curb sets may include too many strategies and minimal curb sets do not provide a useful generalization of the strict equilibrium concept.

Kalai and Samet's (1984) concept of persistent retracts doesn't suffer from the two drawbacks mentioned above. Roughly, this concepts results when requirement (4.5) is weakened to "$B(s) \cap C \neq \emptyset$ for any $s \in U$". Formally, define a *retract* R as a Cartesian product $R = X_i R_i$ where each R_i is a nonempty, closed, convex subset of S_i. A retract is said to be *absorbing* if

$$\mathcal{B}(s) \cap R \neq \emptyset \text{ for all } s \text{ in a neighbourhood } U \text{ of } R, \tag{4.5}$$

that is, if against any small perturbation of strategy profile in R there exists a best response that is in R. A retract is defined to be *persistent* if it is a minimal absorbing retract. Zorn's lemma implies that persistent retracts exist; an elementary proof is indicated below. Kakutani's fixed point theorem implies that each absorbing retract contains a Nash equilibrium. A Nash equilibrium that belongs to a persistent retract is called a *persistent equilibrium*. A slight modification of Myerson's proof for the existence of proper equilibrium actually shows that each absorbing retract contains a proper

equilibrium. Hence, each game has an equilibrium that is both proper and persistent. Below we give examples to show that a proper equilibrium need not be persistent and that a persistent equilibrium need not be proper.

Note that each strict equilibrium is a singleton persistent retract. The reader can easily verify that in the battle-of-the-sexes game only the pure equilibria are persistent and that (in the normal form of) the overall game in Figure 5 only the equilibrium (ps_1, w_2) is persistent, hence, in this example, persistency selects the forward induction outcome. As a side remark, note that $\bar{s}$ is a Nash equilibrium if and only if $\{\bar{s}\} = R$ is a minimal retract with the property "$\mathcal{B}(s) \cap R \neq \emptyset$ for all $s \in R$", hence, persistency corresponds to adding neighborhood robustness to the Nash equilibrium requirement.

Kalai and Samet (1984) show that persistent retracts have a very simple structure, viz. they contain at most one representative from each i-equivalence class of strategies for each player i. To establish this result, Kalai and Samet first note that two strategies s_i' and s_i'' are i-equivalent if and only if there exists an open set U in S such that, against any strategy in U, s_i' and s_i'' are equally good. Hence, it follows that, up to equivalence, the best response of a player is unique (and pure) on an open and dense subset of S. Note that, to a certain extent, a strategy that is not a best response against an open set of beliefs is superfluous, i.e. a player always has a best response that is also a best response to an open set in the neighborhood. Let us call s_i' a *robust best response* against s if there exists an open set $U \in S$ with s in its closure such that s_i' is a best response against all elements in U. (Balkenborg (1992) uses the term semi-robust best response, in order to avoid confusion with Okada's (1983) concept.) Write $\mathcal{B}_i^r(s)$ for all robust best responses of player i against s and $\mathcal{B}^r(s) = X_i \mathcal{B}_i^r(s)$. Note that $\mathcal{B}^r(s) \subset \mathcal{B}^a(s) \subset \mathcal{B}(s)$ for all s. Also note that a mixed strategy is a robust best response only if it is a mixture of equivalent pure robust best responses. Hence, up to equivalence, robustness restricts players to using pure strategies. Finally, note that an outside observer, who is somewhat uncertain about the players' beliefs and who represents this uncertainty by continuous distributions on S, will assign positive probability only to players playing robust best responses.

The reader can easily verify that (4.6) is equivalent to

$$\text{if } s \in R \text{ and } a \in \mathcal{B}_i^r(s) \text{ then } a \sim_i s_i' \text{ for some } s_i' \in R_i \text{ (all } i, s). \tag{4.6}$$

Hence, up to equivalence, all robust best responses against the retract must belong to the absorbing retract. Minimality thus implies that a persistent retract contains at most one representative from each equivalence class of robust best responses. From this observation it follows that there exists an absorbing retract that is spanned by pure strategies and that there exists at least one persistent retract. (Consider the set of all retracts that are spanned by pure strategies. The set is finite, partially ordered and the maximal element ($R = S$) is absorbing, hence, there exists a minimal element.) Of course, for generic strategic form games, no two pure strategies are equivalent and any pure best response is a robust best response. For such games it thus follows that R is a persistent retract if and only if there exists a minimal curb set C such that $R_i = \Delta(C_i)$ for each player i.

We will now investigate which properties from Section 4.2 are satisfied by persistent retracts. We have already seen that persistent retracts exist; they are connected and contain a proper equilibrium. Hence, the properties (E), (C), and (BI) hold. Also (IIA) is satisfied, as follows easily from (4.7) and the fact that $\mathcal{B}^r(s) \subset \mathcal{B}^a(s)$. Also (BRI) follows easily from (4.7). However, persistent retracts do not satisfy (NE). For example, in the matching pennies game the entire set of strategies is the unique persistent retract. Of course, persistency satisfies a weak form of (NE): any persistent retract contains a Nash equilibrium. In fact, it can be shown that each persistent retract contains a stable set of equilibria. (This is easily seen for stability as defined by Kohlberg and Mertens, Mertens (1990) proves it for M-stability and Balkenborg (1992) proves the property for H-stable sets.) Similarly, persistency satisfies a weak form of (A): (4.7) implies that if R is a persistent retract and s_i is an extreme point of R_i, then s_i is a robust best response, hence, s_i is admissible. Consequently, property (A) holds for the extreme points of R, and each element in R only assigns positive probability to admissible pure strategies. This, however, does not imply that the elements of R are themselves admissible. For example, in the game of Figure 9, the only persistent retract is the entire game, but the strategy $(\frac{1}{2}, \frac{1}{2}, 0)$ of player 1 is dominated. In particular the equilibrium

$\langle(\frac{1}{2},\frac{1}{2},0),(0,0,1)\rangle$ is persistent but not perfect.

$3,0$	$0,3$	$0,2$
$0,3$	$3,0$	$0,2$
$2,0$	$2,0$	$0,0$

Figure 9: A persistent equilibrium need not be perfect

Persistent retracts are not invariant. In Figure 9, replace the payoff "2" by "$\frac{3}{2}$" so that the third strategy becomes a duplicate of the mixture $(\frac{1}{2},\frac{1}{2},0)$. The unique persistent retract contains the mixed strategy $(\frac{1}{2},\frac{1}{2},0)$, but it does not contain the equivalent strategy $(0,0,1)$. Hence, the invariance requirement (I) is violated. Balkenborg (1992), however, shows that the extreme points of a persistent retract satisfy (I). He also shows that this set of extreme points satisfies the small worlds property (SWP) and the decomposition property (D).

A serious drawback of persistency is that it does not satisfy the player splitting property: The agent normal form and the normal form of an incomplete information game can have different persistent retracts. The reason is that the normal form forces different types to have the same beliefs about the opponent, whereas the Selten form (i.e. the agent normal form) allows different types to have different conjectures. (Cf. our discussion in Section 2.2.) Perhaps it is even more serious that also other completely inessential changes in the game may induce changes in the persistent retracts and may make equilibria persistent that were not persistent before. As an example, consider the game from Figure 5 in which only the outcome $(3,1)$ is persistent. Now change the game such that, when (pw_1,w_2) is played, the players don't receive zero right away, but are rather forced to play a matching pennies game. Assume players simultaneously choose "heads" or "tails", that player 1 receives 4 units from player 2 if choices match and that he has to pay 4 units if choices differ. The change is completely inessential (the game that was added has unique optimal strategies and value zero), but it has the consequence that in the normal form, only the entire strategy space is persistent. In particular, player

1 taking up his outside option is a persistent and proper equilibrium outcome of the modified game.

For applications of persistent equilibria the reader is referred to Kalai and Samet (1985), Hurkens (1996), Van Damme and Hurkens (1996), Blume (1994, 1996), and Balkenborg (1993). Kalai and Samet consider "repeated" unanimity games. In each of finitely many periods, players simultaneously announce an outcome. The game stops as soon as players announce that same outcome, and then that outcome is implemented. Kalai and Samet show that if there are at least as many rounds as there are outcomes, players will agree on an efficient outcome in a (symmetric) persistent equilibrium. Hurkens (1996) analyzes situations in which some players can publicly burn utility before the play of a game. He shows that if the players who have this option have common interests (Aumann and Sorin (1989)), then only the outcome that these players prefer most is persistent. Van Damme and Hurkens (1996) study games in which players have common interests and in which the timing of the moves is endogenous. They show that persistency forces players to coordinate on the efficient equilibrium. Blume (1994, 1996) applies persistency to a class of signalling games and he also obtains that persistent equilibria have to be efficient. Balkenborg (1993) studies finitely repeated common interest games. He shows that persistent equilibria are almost efficient.

The picture that emerges from these applications (as well as from some theoretical considerations not discussed here, see Van Damme (1992)) is that persistency might be more relevant in an evolutionary and/or learning context, rather than in the pure deductive context we have assumed in this chapter. Indeed, Hurkens (1994) discusses an explicit learning model in which play eventually settles down in a persistent retract. The following proposition summarizes the main elements from the discussion in this section:

Theorem 12 *(i) (Kalai and Samet (1985)). Every game has a persistent retract. Each persistent retract contains a proper equilibrium. Each strategy in a persistent retract assigns positive probability only to robust best replies.*

(ii) (Balkenborg (1992)). For generic strategic form games, persistent retracts correspond to minimal curb sets.

(iii) (Balkenborg (1992)). Persistent retracts satisfy the properties (E), (C), (IIA), (BRI) and (BI) from Section 4.2, but violate the other properties. The set of extreme points of persistent retracts satisfies (SWP), (D) and (I).

(iv) (Mertens (1990), Balkenborg (1992)). Each persistent retract contains an M-stable set. It also contains an H-stable set as well as a KM-stable set.

5 Equilibrium selection

Up to now this paper has been concerned just with the first and basic question of noncooperative game theory: Which outcomes are self-enforcing? The starting point of our investigations was that being a Nash equilibrium is necessary but not sufficient for self-enforcingness, and we have reviewed several other necessary requirements that have been proposed. We have seen that frequently even the most stringent refinements of the Nash concept allow multiple outcomes. For example, many games admit multiple strict equilibria and any such equilibrium passes every test of self-enforcingness that has been proposed up to now. In the introduction, however, we already argued that the "theory" rationale of Nash equilibrium relies essentially on the assumption that players can coordinate on a single outcome. Hence, we have to address the questions of when, why and how players can reach such a coordinated outcome. One way in which such coordination might be achieved is if there exists a convincing theory of rationality that selects a unique outcome in every game and if this theory is common knowledge among the players. One such theory of equilibrium selection has been proposed in Harsanyi and Selten (1988). In this section we will review the main building blocks of that theory.

The theory from Harsanyi and Selten may be seen as derived from three basic postulates, viz. that a theory of rationality should make a recommendation that is (i) a unique strategy profile, (ii) self-enforcing, and (iii) universally applicable. The latter requirement says that no matter the context in which the game arises, the theory should apply. It is a strong form of history-independence. Harsanyi and Selten (1988, pp. 342-43) refer to it as the assumption of endogenous expectations: the solution of the game should depend only on the mathematical structure of the game itself, no matter

the context in which this structure arises. The combination of these postulates is very powerful; for example, one implication is that the solution of a symmetric game should be symmetric. The postulates also force an agent normal form perspective: once a subgame is reached, only the structure of the subgame is relevant, hence, the solution of a game has to project onto the solution of the subgame. Harsanyi and Selten refer to this requirement as "subgame consistency". It is a strong form of the requirement of "persistent rationality" that was extensively discussed in Section 3. Of course, subgame consistency is naturally accompanied by the axiom of truncation consistency: to find the overall solution of the game it should be possible to replace a subgame by its solution. Indeed, Harsanyi and Selten insist on truncation consistency as well. It should now be obvious that the requirements that Harsanyi and Selten impose are very different from the requirements that we discussed in Section 4.2. Indeed the requirements are incompatible. For example, the Harsanyi/Selten requirements imply that the solution of the game from Figure 5 is (tm_1, m_2) where $m_i = \frac{1}{4}s_i + \frac{3}{4}w_i$. Symmetry requires the solution of the subgame to be (m_1, m_2) and the axioms of subgame and truncation consistency prevent player 1 from signalling anything. If one accepts the Harsanyi/Selten postulates, then it is common knowledge that the battle-of-the-sexes subgame has to be played according to the mixed equilibrium, hence, if he has to play, player 2 *must* conclude that player 1 has made a mistake. Note that uniqueness of the solution is already incompatible with the pair (I), (BI) from Section 4.2. We showed that (I) and (BI) leave only the payoff $(3, 1)$ in the game of Figure 5, hence, uniqueness forces $(3, 1)$ as the unique solution of the "battle of the sexes". However, if we would have given the outside option to player 2 rather than to player 1, we would have obtained $(1, 3)$ as the unique solution. Hence, to guarantee existence, the approach from Section 4 must give up uniqueness, i.e. it has to allow multiple solutions. Both (3,1) and (1,3) have to be admitted as solutions of the battle-of-the-sexes game, in order to allow the context in which the game is played to determine which of these equilibria will be selected. The approach to be discussed in this section, which requires context independence, is in sharp conflict with that from the previous section. However, let us note that, although the two approaches are incompatible, each of the approaches corresponds to a coherent point

of view. We confine ourselves to presenting both points of view, to allow the reader to make up his own mind.

5.1 Overview of the Harsanyi/Selten solution procedure

The procedure proposed by Harsanyi and Selten to find the solution of a given game generates a number of "smaller" games which have to be solved by the same procedure. The process of reduction and elimination should continue until finally a basic game is reached which cannot be scaled down any further. The solution of such a basic game can be determined by applying the tracing procedure to which we will return below. Hence, the theory consists of a process of reducing a game to a collection of basic games, a rule for solving each basic game, and a procedure for aggregating these basic solutions to a solution of the overall game. The solution process may be said to consist of five main steps, viz. (i) initialization, (ii) decomposition, (iii) reduction, (iv) formation splitting, and (v) solution using dominance criteria. To describe these steps in somewhat greater detail, we first introduce some terminology.

The Harsanyi/Selten theory makes use of the so-called standard form of a game, a form that is in between the extensive form and the normal form. Formally, the standard form consists of the agent normal form together with information about which agents belong to the same player. Write I for the set of players in the game and for each $i \in I$, let $H_i = \{ij : j \in J_i\}$ be the set of agents of player i. Writing $H = \cup_i H_i$ for the set of all agents in the game, a *game in standard form* is a tuple $g = \langle A, u \rangle_H$ where $A = X_{ij} A_{ij}$ with A_{ij} being the action set of agent ij, and $u_i : A \to \mathsf{R}$ for each player i. Harsanyi and Selten work with this form since on the one hand they want to guarantee perfectness in the extensive form, while on the other hand they want different agents of the same player to have the same expectations about the opponents.

Given a game in extensive form, the Harsanyi/Selten theory should not be directly applied to its associated standard form g; rather, for each $\varepsilon > 0$ that is sufficiently small, the theory should be applied to the uniform ε-perturbation g^ε of the game. The solution of g is obtained by taking the limit, as ε tends to zero, of the solution of g^ε. The question of whether the limit exists is not treated in Harsanyi and Selten (1988);

the authors refer to the unpublished working paper Harsanyi and Selten (1977) in which it is suggested that there should be no difficulties. Formally, g^ε is defined as follows. For each agent ij let σ_{ij} be the *centroid* of A_{ij}, i.e. the strategy that chooses all pure actions in A_{ij} with probability $|A_{ij}|^{-1}$. For $\varepsilon > 0$ sufficiently small, write $\varepsilon_{ij} = \varepsilon|A_{ij}|$ and let $\tilde{\varepsilon} = (\varepsilon_{ij})_{ij \in H}$. Recall, from Equation (3.2) that $s^{\tilde{\varepsilon},\sigma}$ denotes the strategy vector that results when each player intends to play s, but players make mistakes with probabilities determined by $\tilde{\varepsilon}$ and mistakes are given by σ. The *uniformly perturbed game* g^ε is the standard form game $\langle A, u^\varepsilon \rangle_H$ where the payoff function u^ε is defined by $u_i^\varepsilon(s) = u_i(s^{\tilde{\varepsilon},\sigma})$. Hence, in g^ε each agent ij mistakenly chooses each action with probability ε and the total probability that agent ij makes a mistake is $|A_{ij}|\varepsilon$.

Let C be a collection of agents in a standard form game g and denote the complement of C by $\bar{C}$. Given a strategy vector t for the agents in $\bar{C}$, write $g_C^t = \langle A, u^t \rangle_C$ for the *reduced game* faced by the agents in C when the agents in $\bar{C}$ play t, hence $u_{ij}^t(s) = u_{ij}(s,t)$ for $ij \in C$. Write $g_C^t = g_C$ and $u^t = u^C$ in the special case where t is the centroid strategy for each agent in $\bar{C}$. The set C is called *cell* in g if for each t and each player i with an agent in C there exist constants $\alpha_i(t) > 0$ and $\beta_i(t) \in \mathsf{R}$ such that

$$u_i^t(s) = \alpha_i(t) u_i^C(s) + \beta_i(t) \quad \text{(for all } s\text{)}. \tag{5.1}$$

Hence, if C is a cell, then up to positive linear transformations, the payoffs to agents in C are completely determined by the agents in C. Since the intersection of two cells is again a cell whenever this intersection is nonempty, there exist minimal cells. Such cells are called *elementary cells*. Two elementary cells have an empty intersection. Note that for the special case of a normal form game (each player has only one agent), each cell is a small world. Also note that a transformation as in (5.1) leaves the best-reply structure unchanged. Hence, if we had defined a small world as a set of players whose (admissible) best responses are not influenced by outsiders, then each small world would have been a cell. A solution concept that assigns to each standard form game g a unique strategy vector $f(g)$ is said to satisfy *cell and truncation consistency* if for each C that

is a cell in g we have

$$f_{ij}(g) = \begin{cases} f_{ij}(g^C) & \text{if} \quad ij \in C \\ \\ f_{ij}\left(g_{\bar{C}}^{f(g^C)}\right) & \text{if} \quad ij \notin C. \end{cases} \tag{5.2}$$

The reader may check that a subgame of a uniformly perturbed extensive form game induces a cell in the associated perturbed standard form; hence, the axiom of cell and truncation consistency formalizes the idea that the solution is determined by backward induction in the extensive form.

If g is a standard form game and B_{ij} is a nonempty set of actions for each agent ij, then $B = X_{ij}B_{ij}$ is called a *formation* if for each agent ij, each best response against any correlated strategy that only puts probability on actions in B belongs to B_{ij}. Hence, in normal form games, formations are just like curb sets (cf. Section 2.2), the only difference being that formations allow for correlated beliefs. As the intersection of two formations is again a formation, we can speak about *primitive formations*, i.e. formations that do not contain a proper subformation.

An action a of an agent ij is said to be *inferior* if there exists another action b of this agent that is a best reply against a strictly larger set of (possibly) correlated beliefs of the agents. Hence, noninferiority corresponds to the concept of robustness that we encountered (for the case of independent beliefs) in Section 4.5. Any strategy that is weakly dominated is inferior, but the converse need not hold.

Using the concepts introduced above, we can now describe the main steps employed in the Harsanyi/Selten solution procedure:

1. *Initialization*: Form the standard normal form g of the game, and, for each $\varepsilon > 0$ that is sufficiently small, compute the uniformly perturbed game g^ε; compute the solution $f(g^\varepsilon)$ according to the steps described below and put $f(g) = \lim_{\varepsilon \downarrow 0} f(g^\varepsilon)$.

2. *Decomposition*: Decompose the game into its elementary cells; compute the solution of an indecomposable game according to the steps described below and form the solution of the overall game by using cell and truncation consistency.

3. *Reduction*: Reduce the game by using the next three operations:.

   (i) Eliminate all inferior actions of all agents.

   (ii) Replace each set of equivalent actions of each agent ij (i.e. actions among which all players are indifferent) by the centroid of that set.

   (iii) Replace, for each agent ij, each set of ij-equivalent actions (i.e. actions among which ij is indifferent no matter what the others do) by the centroid strategy of that set.

   By applying these steps, an irreducible game results. The solution of such a game is by means of Step 4.

4. *Solution*:

   (i) *Initialization*: Split the game into its primitive formations and determine the solution of each basic game associated with each primitive formation by applying the tracing procedure to the centroid of that formation. The set of all these solutions constitutes the first candidate set Ω^1.

   (ii) *Candidate elimination and substitution*: Given a candidate set Ω, determine the set $M(\Omega)$ of maximally stable elements in Ω. These are those equilibria in Ω that are least dominated in Ω. Dominance involves both payoff dominance and risk dominance and payoff dominance ranks more important than risk dominance. The latter is defined by means of the tracing procedure (see below) and need not be transitive. Form the chain $\Omega = \Omega^1, \Omega^{t+1} = M(\Omega^t)$ until $\Omega^{T+1} = \Omega^T$. If $|\Omega^T| = 1$, then Ω^T is the solution, otherwise replace Ω^T by the trace, $t(\Omega^T)$, of its centroid and repeat the process with the new candidate set $\Omega = \Omega^{T-1} \backslash \Omega^T \cup \{t(\Omega^T)\}$.

It should be noted that it may be necessary to go through these steps repeatedly. Furthermore, the steps are hierarchically ordered, i.e. if the application of Step 3(i) (i.e. the elimination of inferior actions) results in a decomposable game, one should first return to Step 2. The reader is referred to the flow chart on p. 127 of Harsanyi and Selten (1988) for further details.

The next two sections of the present paper are devoted to Step 4, the core of the solution procedure. We conclude this subsection with some remarks on the other steps.
We already discussed Step 2, as well as the reliance on the agent normal form in the previous subsection. Deriving the solution of an unperturbed game as a limit of solutions of uniformly perturbed games has several consequences that might be considered undesirable. For one, duplicating strategies in the unperturbed game may have an effect on the outcome. Consider the normal form of the game from Figure 6. If we duplicate the strategy pl_1 of player 1, the limit solution prescribes r_2 for player 2 (since the mistake pl_1 is more likely than the mistake pr_1), but if we duplicate pr_1 then the solution prescribes player 2 to choose l_2. Hence, the Harsanyi/Selten solution does not satisfy the invariance requirement (I) from Section 4.2, nor does it satisfy (IIA). Secondly, an action that is dominated in the unperturbed game need no longer be dominated in the ε-perturbed version of the game and, consequently, it is possible to construct an example in which the Harsanyi/Selten solution is an equilibrium that uses dominated strategies (Van Damme (1990)). Hence, the Harsanyi/Selten solution violates (A). Turning now to the reduction step, we note that the elimination procedure implies that invariance is violated. (Cf. the discussion on persistency in Section 4.5; note that any pure strategy that is a mixture of non-equivalent pure strategies is inferior.) Let us also remark that stable sets need not survive when an inferior strategy is eliminated. (See Van Damme (1987a, Figure 10.3.1) for an example.) Finally, we note that since the Harsanyi/Selten theory makes use of payoff comparisons of equilibria, the solution of that theory is not best reply invariant. We return to this below.

5.2 Risk dominance in 2×2 games

The core of the Harsanyi/Selten theory of equilibrium selection consists of a procedure that selects, in each situation in which it is common knowledge among the players that there are only two viable solution candidates, one of these candidates as the actual solution for that situation. A simple example of a game with two obvious solution candidates (viz. the strict equilibria (a, a) and $(\bar{a}, \bar{a})$) is the stag-hunt game of the left-hand panel of Figure 10, which is a slight modification of a game first discussed in Aumann

(1990). (The only reason to discuss this variant is to be able to draw simpler pictures).

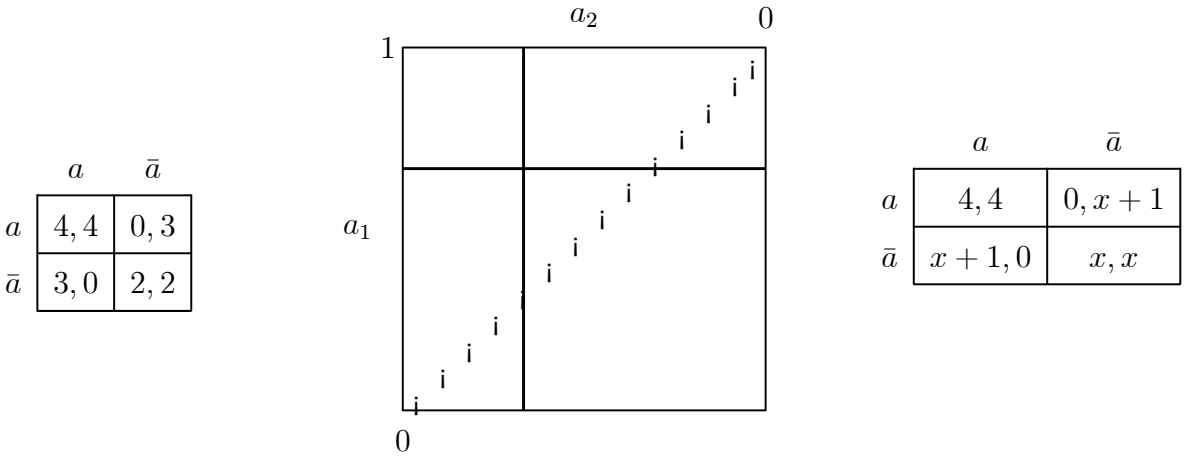


Figure 10: The stag hunt

The stag hunt from the left-hand panel is a symmetric game with common interests (Aumann and Sorin (1989)), i.e. it has (a, a) as the unique Pareto-efficient outcome. Playing a, however, is quite risky: If the opponent plays his alternative equilibrium strategy $\bar{a}$, the payoff is only zero. Playing $\bar{a}$ is much safer: one is guaranteed the equilibrium payoff and, if the opponent deviates, the payoff is even higher. Harsanyi and Selten discuss a variant of this game extensively since it is a case where the two selection criteria that are used in their theory (viz. those of payoff dominance and risk dominance) point in opposite directions. (See Harsanyi and Selten (1988, pp. 88-89, and 358-359).) Obviously, if each player could trust the other to play a, he would also play a, and players clearly prefer such mutual trust to exist. The question, however, is under which conditions such trust exists and how it can be created if it does not exist. As Aumann (1990) has argued, preplay communication cannot create trust where it does not exist initially. In the end, Harsanyi and Selten decide to give precedence to the payoff dominance criterion, i.e. they assume that rational players can rely on collective rationality and they select (a, a) in the game of Figure 10. However, the arguments given are not fully convincing. We will use the game of Figure 10 to illustrate the concept of risk dominance, which is based on strictly individualistic rationality considerations.

Intuitively, the equilibrium s risk dominates the equilibrium $\bar{s}$ if, when players are in a state of mind where they think that either s or $\bar{s}$ should be played, they eventually come to the conclusion that $\bar{s}$ is too risky and, hence, they should play s. For general games, risk dominance is defined by means of the tracing procedure. For the special case of 2-player 2×2 normal form games with two strict equilibria, the concept is also given an axiomatic foundation. Before discussing this axiomatization, we first illustrate how riskiness of an equilibrium can be measured in 2×2 games.

Let $G(a, \bar{a})$ be the set of all 2-player normal form games in which each player i has the strategy set $\{a, \bar{a}\}$ available and in which (a, a) and $(\bar{a}, \bar{a})$ are strict Nash equilibria. For $g \in G(a, \bar{a})$, we identify a mixed strategy of player i with the probability a_i that this strategy assigns to a and we write $\bar{a}_i = 1 - a_i$. We also write $d_i(a)$ for the loss that player i incurs when he unilaterally deviates from (a, a) (hence, $d_1(a) = u_1(a, a) - u_1(\bar{a}, a)$) and we define $d_i(\bar{a})$ similarly. Note that when player j plays a with probability a_j^* given by

$$a_j^* = d_i(\bar{a})/(d_i(a) + d_i(\bar{a})), \tag{5.3}$$

player i is indifferent between a and $\bar{a}$. Hence, the probability a_j^* as in (5.3) represents the risk that i is willing to take at $(\bar{a}, \bar{a})$) before he finds it optimal to switch to a. In a symmetric game (such as that of Figure 10) $a_1^* = a_2^*$, hence a_1^* (resp. $\bar{a}_1^*$) is a natural measure of the riskiness of the equilibrium $(\bar{a}, \bar{a})$ (resp. (a, a)) and $(\bar{a}, \bar{a})$ is more risky if $a_1^* < \bar{a}_1^*$, that is, if $a_1^* < \frac{1}{2}$. In the game of Figure 10, we have that $a_1^* = \frac{2}{3}$, hence (a, a) is more risky than $(\bar{a}, \bar{a})$. More generally, let us measure the riskiness of an equilibrium as the sum of the players' risks. Formally, say that (a, a) *risk dominates* $(\bar{a}, \bar{a})$ in g (abbreviated $a \succ_g \bar{a}$) if

$$a_1^* + a_2^* < 1; \tag{5.4}$$

say that $(\bar{a}, \bar{a})$ risk dominates (a, a) (written $\bar{a} \succ_g a$) if the reverse strict inequality holds, and say that there is no dominance relationship between (a, a) and $(\bar{a}, \bar{a})$) (written $a \sim_g \bar{a}$) if (5.4) holds with equality. In the game of Figure 10, we have that $(\bar{a}, \bar{a})$ risk dominates (a, a).

To show that these definitions are not "ad hoc", we now give an axiomatization of risk-dominance. On the class $G(a, \bar{a})$, Harsanyi and Selten (1988, Section 3.9) characterize this relation by the following axioms.

1. (Asymmetry and completeness): For each g exactly one of the following holds: $a \succ_g \bar{a}$ or $\bar{a} \succ_g a$ or $a \sim_g \bar{a}$.

2. (Symmetry): If g is symmetric and player i prefers (a, a) while player $j (j \neq i)$ prefers $(\bar{a}, \bar{a})$, then $a \sim_g \bar{a}$.

3. (Best-reply invariance): If g and g' have the same best-reply correspondence, then $a \succ_g \bar{a}$, if and only if $a \succ_{g'} \bar{a}$.

4. (Payoff monotonicity): If g' results from g by making (a, a) more attractive for some player i while keeping all other payoffs the same, then $a \succ_{g'} \bar{a}$ whenever $a \succ_g \bar{a}$ or $a \sim_g \bar{a}$.

The proof is simple and follows from the observations that

(i) games are best-reply-equivalent if and only if they have the same (a_1^*, a_2^*),

(ii) symmetric games with conflicting interests satisfy (5.4) with equality, and

(iii) increasing $u_i(a, a)$ decreases a_j^*.

Harsanyi/Selten also give an alternative characterization of risk-dominance. Condition (5.4) is equivalent to the (Nash) product of players' deviation losses at (a, a) being larger than the corresponding Nash product at $(\bar{a}, \bar{a})$, hence

$$d_1(a)d_2(a) > d_1(\bar{a})d_2(\bar{a}) \tag{5.5}$$

and, in fact, the original definition is by means of this inequality. Yet another equivalent characterization is that the area of the stability region of (a, a) (i.e. the set of mixed strategies against which a is a best response for each player) is larger than the area of the stability region of $(\bar{a}, \bar{a})$. (Obviously, the first area is $\bar{a}_1^* \bar{a}_2^*$, the second is $a_1^* a_2^*$.) For the stag hunt game, the stability regions have been displayed in the middle panel of Figure 10. (The diagonal represents the line $a_1 + a_2 = 1$; the upper left corner of the diagram is the point $a_1 = 1$, $a_2 = 1$, it corresponds to the upper left corner of the matrix, and similarly for other points.)

In Carlsson and Van Damme (1993a), equilibrium selection according to the risk-dominance criterion is derived from considerations related to uncertainty concerning the payoffs of the game. These authors assume that players can observe the payoffs in a game only with some noise. In contrast to Harsanyi's model that was discussed in Section 2.5, Carlsson and Van Damme assume that each player is uncertain about both players' payoffs. Because of the noise, the actual best-reply structure will not be common knowledge and as a consequence of this lack of common knowledge, players' behavior at each observation may be governed by the behavior at some remote observation (also cf. Rubinstein (1989)). In the noisy version of the stag hunt game of Figure 8, even though players may know to a very high degree that (a, a) is the Pareto-dominant equilibrium, they might be unwilling to play it since each player i might think that j will play $\bar{a}$ since i will think that j will think ... that $\bar{a}$ is a dominant action. Hence, even though this model superficially resembles that of Harsanyi (1973a), it leads to completely different results.

As a simple and concrete illustration of the model, suppose that it is common knowledge among the players that payoffs are related to actions as in the right panel $g(x)$ of Figure 10. A priori, players consider all values $x \in [-1, 4]$ to be possible and they consider all such values to be equally likely. (Carlsson and Van Damme (1993a) show that the conclusion is robust with respect to such distributional assumptions, as well as with respect to assumptions on the structure of the noise). Note that $g(x) \in G(a, \bar{a})$ for $x \in (0, 3)$, that a is a dominant strategy if $x < 0$ and that $\bar{a}$ is dominant if $x > 3$. Suppose now that players can observe the actual value of x that prevails only with some slight noise. Specifically, assume player i observes $x_i = x + \varepsilon e_i$ where x, e_1, e_2 are independent and e_i is uniformly distributed on $[-1, 1]$. Obviously, if $x_i < -\varepsilon$ (resp. $x_i > 3 + \varepsilon$), player i will play a (resp. $\bar{a}$) since he knows that that action is dominant at each actual value of x that corresponds to such an observation. Forcing players to play their dominant actions at these observations will make a and $\bar{a}$ dominant at a larger set of observations and the process can be continued iteratively. Let $\underline{x}$ (resp. $\bar{x}$) be the supremum (resp. infimum) of the set of observations y for which each player i has a (resp. $\bar{a}$) as an iteratively dominant action for each $x_i < y$ (resp. $x_i > y$). Then there

must be a player i who is indifferent between a and $\bar{a}$ when he observes $\underline{x}$ (resp. $\bar{x}$). Writing $a_j(x_i)$ for the probability that i assigns to j playing a when he observes x_i, we can write the indifference condition of player i at x_i (approximately) as

$$4a_j(x_i) = a_j(x_i) + x_i. \tag{5.6}$$

Now, at $x_i = \underline{x}$, we have that $a_j(x_i)$ is at least $\frac{1}{2}$ because of our symmetry assumptions and since j has a as an iteratively dominant strategy for each $x_j < \underline{x}$. Consequently, $\underline{x} \geq \frac{3}{2}$. A symmetric argument establishes that $\bar{x} \leq \frac{3}{2}$, hence $\underline{x} = \bar{x} = \frac{3}{2}$, and each player i should choose a if he observes $x_i < \frac{3}{2}$ while he should choose $\bar{a}$ if $x_i > \frac{3}{2}$. Hence, in the noisy version of the game, each player should always play the risk-dominant equilibrium of the game that corresponds to his observation.

To conclude this subsection, we remark that the concept of risk dominance also plays an important role in the literature that derives Nash equilibrium as a stationary state of processes of learning or evolution. Even though each Nash equilibrium may be a stationary state of such a process, occasional experimentation or mutation may result in only the risk-dominant equilibrium surviving in the long run: This equilibrium has a larger stability region, hence, a larger basin of attraction, so that the process is more easily trapped there and mutations have more difficulty to upset it (See Kandori et al. (1993), Young (1993a,b), Ellison (1993)).

5.3 Risk dominance and the tracing procedure

Let us now consider a more general normal form game $g = \langle A, u \rangle$ where the players are uncertain which of two equilibria, s or $\bar{s}$, should be played. Risk dominance tries to capture the idea that in this state of confusion the players enter a process of expectation formation that converges on that equilibrium which is the least risky of the two. (Note that a player i with $s_i = \bar{s}_i$ is not confused at all. Harsanyi and Selten first eliminate all such players before making risk comparisons. For the remaining players they similarly delete strategies not in the formation spanned by s and $\bar{s}$ since these are never best responses, no matter what expectations the players have. To the smaller game that results in this way, one should then first apply the decomposition and reduction steps

from Section 5.2. We'll assume that all these transformations have been made and we will denote the resulting game again by g.)

Harsanyi and Selten view the rational formation of expectations as a two-stage process. In the first stage, players form preliminary expectations which are based on the structure of the game. These preliminary expectations take the form of a mixed strategy vector s^0 for the game. On the basis of s^0, players can already form plans about how to play the game. A naive plan would be for each player to play the best response against s^0, but, of course, these plans are not necessarily consistent with the preliminary expectations. The second stage of the expectation formation process then consists of a procedure that gradually adjusts plans and expectations until they are consistent and yield an equilibrium of the game g. Harsanyi and Selten actually make use of two adjustment processes, the linear tracing procedure T and the logarithmic tracing procedure $\tilde{T}$. Formally, each of these is a map that assigns to a mixed strategy vector s^0 exactly one equilibrium of g. The linear tracing procedure is easier to work with, but it is not always well-defined. The logarithmic tracing procedure is well-defined and yields the same outcome as the linear one whenever the latter is well-defined. We now first discuss these tracing procedures. Thereafter, we return to the question of how to form the preliminary expectations and how to define risk dominance for general games.

Let $g = \langle A, u\rangle$ be a normal form game and let p be a vector of mixed strategies for g, interpreted as the players' prior expectations. For $t \in [0,1]$ define the game $g^{t,p} = \langle A, u^{t,p}\rangle$ by

$$u_i^{t,p} = tu_i(s) + (1-t)u_i(p\backslash s_i). \tag{5.7}$$

Hence, for $t = 1$ the game coincides with g, while $g^{0,p}$ is a trivial game in which each player's payoff depends only on this player's prior expectations, not on what the opponents are actually doing. Write $\Gamma(p)$ for the graph of the equilibrium correspondence, hence

$$\Gamma(p) = \{(t,s) \in [0,1] \times S : s \text{ is an equilibrium of } g^{t,p}\}. \tag{5.8}$$

In nondegenerate cases, $g^{0,p}$ will have exactly one (and strict) equilibrium $s(0,p)$ and this equilibrium will remain an equilibrium for sufficiently small t. Let us denote it by $s(t,p)$.

The linear tracing procedure now consists in following the curve $s(t,p)$ until, at its endpoint $T(p) = s(1,p)$, an equilibrium of g is reached. Hence, as the tracing procedure progresses, plans and expectations are continuously adjusted until an equilibrium is reached. The parameter t may be interpreted as the degree of confidence players have in the solution $s(t,p)$. Formally, the *linear tracing procedure* with prior p is well-defined if the graph $\Gamma(p)$ contains a unique connected curve that contains endpoints both at $t = 0$ and $t = 1$. In this case, the endpoint $T(p)$ at $t = 1$ is called the *linear trace* of p. (Note the requirement that there be a unique connecting curve. Herings (2000) shows that there will always be at least one such curve, hence, the procedure is feasible in principle.)

We can illustrate the procedure by means of the stag hunt game from Figure 10. Write p_i for the prior probability that i plays a. If $p_i > \frac{2}{3}$ for $i = 1, 2$, then $g^{0,p}$ has (a,a) as its unique equilibrium and this strategy pair remains an equilibrium for all t. Furthermore, for any $t \in [0,1]$, (a,a) is disconnected in $\Gamma(p)$ from any other equilibrium of $g^{t,p}$. Hence, in this case the linear tracing procedure is well-defined and we have $T(p) = (a,a)$. Similarly, $T(p) = (\bar{a},\bar{a})$ if $p_i < \frac{2}{3}$ for $i = 1, 2$. Next, assume $p_1 < \frac{2}{3}$ and $p_2 > \frac{2}{3}$ so that $s(0,p) = (a,\bar{a})$. In this case the initial plans do not constitute an equilibrium of the final game so that adjustments have to take place along the path. The strategy pair $(a,\bar{a})$ remains an equilibrium of $g^{t,p}$ as long as

$$4(1-t)p_2 \geq 2t + (1-t)(2+p_2) \tag{5.9}$$

and

$$(1-t)(2+p_1) + 3t \geq 4p_1(1-t) + 4t. \tag{5.10}$$

Hence, provided that no player switches before t, player 1 has to switch at the value of t given by

$$t/(1-t) = (3p_2 - 2)/2 \tag{5.11}$$

while player 2 has to switch when

$$t/(1-t) = 2 - 3p_1. \tag{5.12}$$

Assume $p_1 + p_2/2 < 1$ so that the t-value determined by (5.11) is smaller than the value determined by (5.12). Hence, player 1 has to switch first and, following the branch $(a, \bar{a})$, the linear tracing procedure continues with a branch $(\bar{a}, \bar{a})$. Since $(\bar{a}, \bar{a})$ is a strict equilibrium of g, this branch continues until $t = 1$, hence $T(p) = (\bar{a}, \bar{a})$ in this case. Similarly, $T(p) = (a, a)$ if $p_1 < \frac{2}{3}, p_2 > \frac{2}{3}$ and $p_1 + p_2/2 > 1$. In the case where $p_1 > \frac{2}{3}$ and $p_2 < \frac{2}{3}$, the linear trace of p follows by symmetry. The results of our computations are summarized in the left-hand panel of Figure 11.

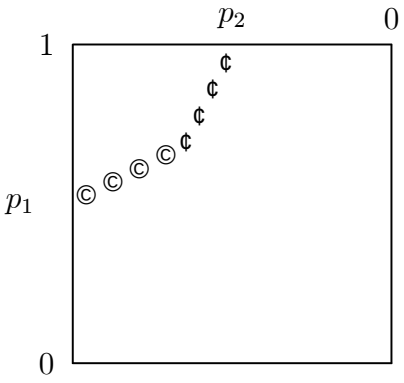


Figure 11a: In the interior of the shaded area $T(p) = (a, a)$. In the interior of the complement $T(p) = (\bar{a}, \bar{a})$.

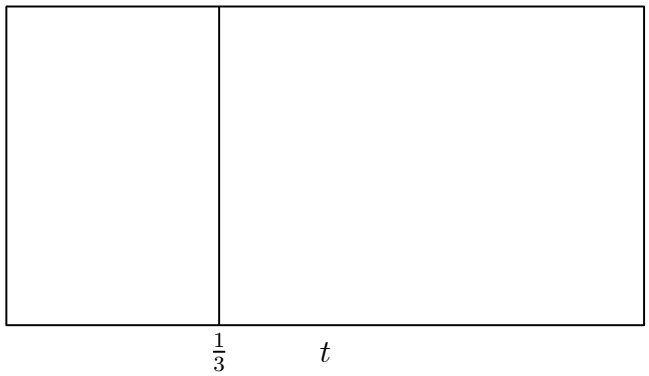


Figure 11b: A case where the linear tracing procedure is not well-defined.

If $p_1 < \frac{2}{3}, p_2 > \frac{2}{3}$ and $p_1 + p_2/2 = 1$, then the equations (5.11)-(5.12) determine the same t-value, hence, both players want to switch at the same time $\tilde{t}$. In this case, the game $g^{\tilde{t},p}$ is degenerate with equilibria both at (a, a) and at $(\bar{a}, \bar{a})$. Now there exists a path in Γ that connects $(a, \bar{a})$ with (a, a) as well as a path that connects $(a, \bar{a})$ with $(\bar{a}, \bar{a})$. In fact, all three equilibria of g (including the mixed one) are connected to the equilibrium of $g^{0,p}$, hence, the linear tracing procedure is not well-defined in this case. Figure 11b gives a graphical display of this case. (The picture is drawn for the case where $p_1 = \frac{1}{2}, p_2 = 1$ and displays the probability of 1 choosing a.)

The logarithmic tracing procedure has been designed to resolve ambiguities such as those in Figure 11b. For $\varepsilon \in (0, 1], t \in [0, 1)$ and $p \in S$, define the game $g^{\varepsilon,t,p}$ by means

of

$$u_i^{\varepsilon,t,p}(s) = u_i^{t,p}(s) + \varepsilon(1-t)\alpha_i \sum_a \ln s_i(a) \tag{5.13}$$

where α_i is a constant defined by

$$\alpha_i = \max_s [\max_{s_i'} u_i(s \backslash s_i') - \min_{s_i'} u_i(s \backslash s_i')]. \tag{5.14}$$

Hence, $u_i^{\varepsilon,t,p}(s)$ results from adding a logarithmic penalty term to $u_i^{t,p}(s)$. This term ensures that all equilibria are completely mixed and that there is a unique equilibrium $s(\varepsilon, 0, p)$ if $t = 0$. Write $\tilde{\Gamma}(p)$ for the graph of the equilibrium correspondence

$$\tilde{\Gamma}(p) = \{(\varepsilon, t, s) \in (0,1] \times [0,1) \times S : s \text{ is an equilibrium of } g^{\varepsilon,t,p}\}. \tag{5.15}$$

$\tilde{\Gamma}(p)$ is the zero set of a polynomial and, hence, is an algebraic set. Loosely speaking, the *logarithmic tracing procedure* consists of following, for each $\varepsilon > 0$, the analytic continuation $s(\varepsilon, t, p)$ of $s(\varepsilon, 0, p)$ till $t = 1$ and then taking the limit, as $\varepsilon \to 0$, of the end points. Harsanyi and Selten (1988) and Harsanyi (1975) claim that this construction can indeed be carried out, but Schanuel et al. (1991) pointed to some difficulties in this construction: The analytic continuation need not be a curve and there is no reason for the limit to exist. Fortunately, these authors also showed that, apart from a finite set E of ε-values, the construction proposed by Harsanyi and Selten is indeed feasible. Specifically, if $\varepsilon \notin E$, then there exists a unique analytic curve in $\tilde{\Gamma}(p)$ that contains $s(\varepsilon, 0, p)$. If we write $s(\varepsilon, t, p)$ for the strategy component of this curve, then $\tilde{T}(p) = \lim_{\varepsilon \downarrow 0} \lim_{t \to 1} s(\varepsilon, t, p)$ exists. $\tilde{T}(p)$ is called the *logarithmic trace* of p. Hence, the logarithmic tracing procedure is well-defined. Furthermore, Schanuel et al. (1991) show that there exists a connected curve in $\Gamma(p)$ connecting $\tilde{T}(p)$ to an equilibrium in $g^{0,p}$ implying that $\tilde{T}(p) = T(p)$ whenever the latter is well-defined. Hence, we have

Theorem 13 *(Harsanyi (1975), Schanuel et al. (1991)). The logarithmic tracing procedure $\tilde{T}$ is well-defined. The linear tracing procedure T is well-defined for almost all priors and $\tilde{T}(p) = T(p)$ whenever the latter is well-defined.*

The logarithmic penalty term occurring in (5.13) gives players an incentive to use completely mixed strategies. It has the consequence that in Figure 11b the interior mixed strategy path is approximated as $\varepsilon \to 0$. Hence, if p is on the south-east boundary of the shaded region in Figure 11a, then $\tilde{T}(p)$ is the mixed strategy equilibrium of the game g. We finally come to the construction of the prior probability distribution p used in the risk dominance comparison between s and $\bar{s}$. According to Harsanyi and Selten, each player i will initially assume that his opponents already know whether s or $\bar{s}$ is the solution. Player i will assign a subjective probability z_i to the solution being s and a probability $\bar{z}_i = 1 - z_i$ to the solution being $\bar{s}$. Given his beliefs z_i player i will then choose a best response $b_i^{z_i}$ to the correlated strategy $z_i s_{-i} + \bar{z}_i \bar{s}_{-i}$ of his opponents. (In case of multiple best responses, i chooses all of them with the same probability.) An opponent j of player i is assumed not to know i's subjective probability z_i; however, j knows that i is following the above reasoning process. Applying the principle of insufficient reasoning, Harsanyi/Selten assume that j considers all values of z_i to be equally likely, hence, j considers z_i to be uniformly distributed on $[0, 1]$. Consequently, j believes that i will play $a_i \in A_i$ with a probability given by

$$p_i(a_i) = \int b_i^{z_i}(a_i) dz_i. \tag{5.16}$$

Equation (5.16) determines the players' *prior expectations* p to be used for risk-dominance comparison between s and $\bar{s}$. If $\tilde{T}(p) = s$ (resp. $\tilde{T}(p) = \bar{s}$) then s is said to *risk dominate* $\bar{s}$ (resp. $\bar{s}$ risk dominates s). If $\tilde{T}(p) \notin \{s, \bar{s}\}$, neither equilibrium risk dominates the other. The reader may verify that for 2×2 games this definition of risk dominance is in agreement with the one given in the previous section. For example, in the stag hunt game from Figure 8 we have that $b_i^{z_i}(a) = 1$ if $z_i > \frac{2}{3}$ and $b_i^{z_i}(a) = 0$ if $z_i < \frac{2}{3}$, hence $p_i(a) = \frac{1}{3}$. Consequently, p lies in the non-shaded region in Figure 11a and $T(p) = (\bar{a}, \bar{a})$, hence, $(\bar{a}, \bar{a})$ risk dominates (a, a).

Unfortunately, for games larger than 2×2, the risk dominance relation need not be transitive (see Harsanyi and Selten (1988, Figure 3.25) for an example) and selection on the basis of this criterion need not be in agreement with selection on the basis of stability with respect to payoff perturbations (Carlsson and Van Damme (1993b)). To

illustrate the latter, consider the n-player stag hunt game in which each player i has the strategy set $\{a, \bar{a}\}$. A player choosing a gets the payoff 1 if all players choose a, and 0 otherwise. A player choosing $\bar{a}$ gets the payoff $x \in (0,1)$ irrespective of what the others do. There are two strict Nash equilibria, viz. "all a" and "all $\bar{a}$". If player i assigns prior probability z to his opponents playing the former, then he will play a if $z > x$, hence, $p_i(a) = 1 - x$ according to (5.16). Consequently, the risk-dominant solution is "all a" if

$$(1-x)^{n-1} > x \tag{5.17}$$

and it is "all $\bar{a}$" if the reverse strict inequality is satisfied. On the other hand, Carlsson and Van Damme (1993b) derive that, whenever there is slight payoff uncertainty, a player should play a if $\frac{1}{n} > x$. It is interesting to note that this n-person stag hunt game has a potential (cf. Section 2.3) and that the solution identified by Carlsson/Van Damme maximizes the potential. More generally, suppose that, when there are k players choosing a, the payoff to a player choosing a equals $f(k)$ (with $f(0) = 0$, $f(n) = 1$) and that the payoff to a player choosing $\bar{a}$ equals $x \in (0,1)$. Then the function p that assigns to each outcome in which exactly k players cooperate the value

$$p(k) = \sum_{l=1}^{k} [f(l) - x] \tag{5.18}$$

is an exact potential for the game. "All a" maximizes the potential if and only if $\sum_{l=1}^{k} f(l)/n > x$ and this condition is identical to the one that Carlsson/Van Damme derive for a to be optimal in their model.

To conclude this subsection, we remark that, in order to derive (5.16), it was assumed that player i's uncertainty can be represented by a correlated strategy of the opponents. Güth (1985) argues that such correlated beliefs may reflect the strategic aspects rather poorly and he gives an example to show that such a correlated belief may lead to counterintuitive results. Güth suggests computing the prior as above, save by starting from the assumption that i believes $j \neq i$ to play $z_j s_j + \bar{z}_j \bar{s}_j$ with z_j uniform on $[0,1]$ and different z's being independent.

5.4 Risk dominance and payoff dominance

We already encountered the fundamental conflict between risk dominance and payoff dominance when discussing the stag hunt game in Section 5.2 (Figure 10). In that game, the equilibrium (a, a) Pareto dominates the equilibrium $(\overline{a}, \overline{a})$, but the latter is risk dominant. In cases of such conflict, Harsanyi/Selten have given precedence to the payoff dominance criterion, but their arguments for doing so are not compelling, as they indeed admit in the postscript of their book, when they discuss Aumann's argument (also already mentioned in Section 5.2) that pre-play communication cannot make a difference in this game. After all, no matter what a player intends to play he will always attempt to induce the other to play a as he always benefits from this. Knowing this, the opponent cannot attach specific meaning to the proposal to play (a, a), communication cannot change a player's beliefs about what the opponent will do and, hence, communication can make no difference to the outcome of the game (Aumann (1990)). As Harsanyi and Selten (1988, p. 359) write "This shows that in general we cannot expect the players to implement payoff dominance unless, from the very beginning, payoff dominance is part of the rationality concept they are using. Free communication among the players in itself might not help. Thus if one feels that payoff dominance is an essential aspect of game-theoretic rationality, then one must explicitly incorporate it into one's concept of rationality".

Several equilibrium concepts exist that explicity incorporate such considerations. The most demanding concept is Aumann's (1959) notion of a *strong equilibrium*: it requires that no coalition can deviate in a way that makes all its members better off. Already in simple examples such as the prisoners' dilemma, this concept generates an empty set of outcomes. (In fact, generically all Nash equilibria are inefficient (see Dubey (1986)).) Less demanding is the idea that the grand coalition not be able to renegotiate to a more attractive stable outcome. This idea underlies the concept of renegotiation-proof equilibrium from the literature on repeated games (see Bernheim and Ray (1989), Farrell and Maskin (1989) and Van Damme (1988, 1989a)). Bernheim et al. (1987) have proposed the interesting concept of *coalition-proof Nash equilibrium* as a formalization of the requirement that no subcoalition should be able to profitably deviate to a strategy

vector that is stable with respect to further renegotiation. The concept is defined for all normal form games and the formal definition is by induction on the number of players. For a one-person game any payoff-maximizing action is defined to be coalition-proof. For an I-person game, a strategy profile s is said to be weakly coalition-proof, if, for any proper subcoalition coalition C of I, the strategy profile s_C is coalition-proof in the reduced game in which the complement $\overline{C}$ is restricted to play $s_{\overline{C}}$, and s is said to be coalition-proof if there is no other weakly-coalition proof profile s' that strictly Pareto dominates it. For 2-player games, coalition-proof equilibria exist, but existence for larger games is not guaranteed. Furthermore, coalition-proof equilibria may be Pareto dominated by other equilibria.

The tension between "global" payoff dominance and "local" efficiency was already pointed out in Harsanyi and Selten (1988): an agreement on a Pareto-efficient equilibrium may not be self-enforcing since, with the agreement in place, and accepting the logic of the concept, a subcoalition may deviate to an even more profitable agreement. The following provides a simple example. Consider the 3-player game g in which player 3 first decides whether to take up an outside option T (which yields all players the payoff 1) or to let players 1 and 2 play a subgame in which the payoffs are as in Figure 12.

	a	$\bar{a}$
a	$2, 2, 2$	$0, 0, 0$
$\bar{a}$	$0, 0, 0$	$3, 3, 0$

Figure 12: Renegotiation as a constraint

The game g from Figure 12 has two Nash equilibrium outcomes. In the first, player 3 chooses T (in the belief that 1 and 2 will choose $\bar{a}$ with sufficiently high probability); in the second, player 3 chooses p, i.e. he gives the move to players 1 and 2, who play (a, a). Both outcomes are subgame perfect (even stable) and the equilibrium (a, a, p) Pareto dominates the equilibrium T. At the beginning of the game it seems in the interest of all players to play (a, a, p). However, once player 3 has made his move, his interests have become strategically irrelevant and it is in the interest of players 1 and 2 to renegotiate to $(\bar{a}, \bar{a})$.

Although the above argument was couched in terms of the extensive form of the game, it is equally relevant for the case in which the game is given in strategic form, i.e. when players have to move simultaneously. After agreeing to play (a, a, p), players 1 and 2 could secretly get together and arrange a joint deviation to $(\bar{a}, \bar{a})$. This deviation is in their interest and it is stable since no further deviations by subgroups are profitable. Hence, the profile (a, a, p) is not coalition-proof.

The reader may argue that these "cooperative refinements" in which coalitions of players are allowed to deviate jointly have no place in the theory of strategic equilibrium, and that, as suggested in Nash (1953), it is preferable to stay squarely within the non-cooperative framework and to fully incorporate possibilities for communication and cooperation in the game rather than in the solution concept. The present author agrees with that view. The above discussion has been included to show that, while it is tempting to argue that equilibria that are Pareto-inferior should be discarded, this view encounters difficulties and may not stand up to closer scrutinity. Nevertheless, the shortcut may sometimes yield valuable insights. The interested reader is referred to Bernheim and Whinston (1987) for some applications using the shortcut of coalition-proofness.

5.5 Applications and variations

Nash (1953) already noted the need for a theory of equilibrium selection for the study of bargaining. He wrote: "Thus the equilibrium points do not lead us immediately to a solution of the game. But if we discriminate between them by studying their relative stabilities we can escape from this troublesome nonuniqueness" (Nash (1953, pp. 131-132)). Nash studied 2-person bargaining games in which the players simultaneously make payoff demands, and in which each player receives his demand if and only if the pair of demands is feasible. Since each pair that is just compatible (i.e. is Pareto optimal) is a strict equilibrium, there are multiple equilibria. Using a perturbation argument, Nash suggested taking that equilibrium in which the product of the utility gains is largest as the solution of the game. The desire to have a solution with this "Nash product property" has been an important guiding principle for Harsanyi and Selten when developing their theory (cf. (5.5)). One of the first applications of that

theory was to unanimity games, i.e. games in which each player's payoff is zero unless all players simultaneously choose the same alternative. As the reader can easily verify, the Harsanyi/Selten solution of such a game is indeed the outcome in which the product of the payoffs is largest, provided that there is such a unique maximizing outcome.

Another early application of the theory was to market entry games (Selten and Güth (1982)). In such a game there are I players who simultaneously decide whether to enter a market or not. If k players enter, the payoff to a player i that enters is $\pi(k) - c_i$, while his payoff is zero otherwise (π is a decreasing function). The Harsanyi/Selten solution prescribes entry of the players with the lowest entry costs up to the point where entry becomes unprofitable.

The Harsanyi/Selten theory has been extensively applied to bargaining problems (cf. Harsanyi and Selten (1988, Chs. 6-9), Harsanyi (1980, 1982), Leopold-Wildenburger (1985), Selten and Güth (1991), Selten and Leopold (1983)). Such problems are modelled as unanimity games, i.e. a set of possible agreements is specified, players simultaneously choose an agreement and an agreement is implemented if and only if it is chosen by all players. In case there is no agreement, trade does not take place. For example, consider bargaining between two risk-neutral players about how to divide one dollar and suppose that one of the players, say player 1, has an outside option of α. The Harsanyi/Selten solution allocates $\max(\sqrt{\alpha}, \frac{1}{2})$ to player 1 and the rest to player 2. Hence, the outside option influences the outcome only if it is sufficiently high (Harsanyi and Selten (1988, Ch. 6)). As another example, consider bargaining between one seller and n identical buyers about the sale of an indivisible object. If the seller's value is 0 and each buyer's value is 1, the Harsanyi/Selten solution is that each player proposes a sale at the price $p(n) = (2^n - 1)/(2^n - 1 + n)$.

Harsanyi and Selten (1988, Chs. 8 and 9) apply the theory to simple bargaining games with incomplete information. Players bargain about how to divide one dollar; if there is disagreement, a player receives his conflict payoff, which may be either 0 or x (both with probability $\frac{1}{2}$)) and which is private information. In the case of one-sided incomplete information (it is common knowledge that player 1's conflict payoff is zero), player 1 proposes that he get a share $x(\alpha)$ of the cake, where $x(\alpha)$ is some decreasing square root

function of α with $x(0) = 50$. The weak type of player 2 (i.e. the one with conflict payoff 0) proposes that player 1 get $x(\alpha)$, while the strong type proposes $x(\alpha)$ if $\alpha < \alpha^*(\approx 81)$ and 0 in case $\alpha > \alpha^*$. Hence, the bargaining outcome may be ex post inefficient. Güth and Selten (1991) consider a simple version of Akerlof's lemons problem (Akerlof (1970)). A seller and a buyer are bargaining about the price of an object of art, which may be either worth 0 to each of them (it is a forgery) or which may be worth 1 to the seller and $v > 1$ to the buyer. The seller knows whether the object is original or fake, but the buyer only knows that both possibilities have positive probability. The solution either is disagreement, or exploitation of the buyer by the seller (i.e. the price equals the buyer's expected value), or some compromise in which the buyer bears a greater part of the fake risk than the seller does. At some parameter values, the solution (the price) changes discontinuously, and Güth/Selten admit that they cannot give plausible intuitive interpretations for these jumps.

Van Damme and Güth (1991a,b) apply the Harsanyi/Selten theory to signalling games. In Van Damme and Güth (1991a) the most simple version of the Spence (1973) signalling game is considered. There are two types of workers, one productive, the other unproductive, who differ in their education costs and who can use the education level to signal their type to uninformed employers who compete in prices à la Bertrand. It turns out that the Harsanyi/Selten solution coincides with the E_2-equilibrium that was proposed in Wilson (1977). Hence, the solution is the sequential equilibrium that is most preferred by the high quality worker, and this worker signals his type if and only if signalling yields higher utility than pooling with the unproductive worker does. It is worth remarking that this solution is obtained without invoking payoff dominance. Note that the solution is again discontinuous in the parameter of the problem, i.e. in the ex ante probability that the worker is productive. The discontinuity arises at points where a different element of the Harsanyi/Selten solution procedure has to be invoked. Specifically, if the probability of the worker being unproductive is small, then there is only one primitive formation and this contains only the Pareto-optimal pooling equilibrium. As soon as this probability exceeds a certain threshold, however, also the formation spanned by the Pareto-optimal separating equilibrium is primitive, and, since the sepa-

rating equilibrium risk dominates the pooling equilibrium, the solution is separating in this case.

We conclude this subsection by mentioning some variations of the Harsanyi/Selten theory that have recently been proposed. Güth and Kalkofen (1989) propose the ESBORA theory, whose main difference to the Harsanyi/Selten theory is that the (intransitive) risk dominance relation is replaced by the transitive relation of resistance dominance. The latter takes the intensity of the dominance relation into account. Formally, given two equilibria s and s', define player i's resistance at s against s' as the largest probability z such that, when each player $j \neq i$ plays $(1-z)s_j + zs'_j$, player i still prefers s_i to s'_i. Güth and Kalkofen propose ways to aggregate these individual resistances into a resistance of s against s' which can be measured by a number $r(s, s')$. The resistance against s' can then be represented by the vector $R(s') = \langle r(s, s') \rangle_s$ and Güth/Kalkofen propose to select that equilibrium s' for which the vector $R(s')$, written in nonincreasing order, is lexicographically minimal. At present the ESBORA theory is still incomplete: The individual resistances can be aggregated in various ways and the solution may depend in an essential way on which aggregation procedure is adopted, as examples in Güth and Kalkofen (1989) show (see also Güth (1992) for different aggregation procedures).

For a restricted class of games (specifically, bipolar games with linear incentives), Selten (1995) proposes a set of axioms that determine a unique rule to aggregate the players individual resistances into an overall measure of resistance (or risk) dominance. For 2×2 games, selection on the basis of this measure is in agreement with selection as in Section 5.2, but for larger games, this need no longer be true. In fact, for 2-player games with incomplete information, selection according to the measure proposed in Selten (1995) has close relations with selection according to the "Generalized Nash product" as in Harsanyi and Selten (1972).

Finally, we mention that Harsanyi (1995) proposes to replace the bilateral risk comparisons between pairs of equilibria by a multilateral comparison involving all equilibria that directly identifies the least risky of all of them. He also proposes not to make use of payoff comparisons, a suggestion that brings us back the to fundamental conflict between payoff dominance and risk dominance that was discussed in Section 5.4.

5.6 Final Remark

We end this section and chapter by mentioning a result from Norde et al. (1996) that puts all the attempts to select a unique equilibrium in a different perspective. Recall that in Section 2 we discussed the axiomatization of Nash equilibrium using the concept of consistency, i.e. the idea that a solution of a game should induce a solution of any reduced game in which some players are committed to play the solution. Norde et al. (1996) show that if s is a Nash equilibrium of a game g, g can be embedded in a larger game that only has s as an equilibrium, consequently consistency is incompatible with equilibrium selection. More precisely, Norde et al. (1996) show that the only solution concept that satisfies consistency, nonemptiness and one-person rationality is the Nash concept itself, so that not only equilibrium selection, but even the attempt to refine the Nash concept is frustrated if one insists on consistency.

References

Akerlof, G. (1970). "The Market for Lemons", *Quarterly Journal of Economics* **84**, 488-500.

Aumann, R.J. (1959). "Acceptable Points in General Cooperative n-Person Games", in: R.D. Luce and A.W. Tucker (eds.), *Contributions to the Theory of Games, IV* (Ann. Math. Study 40), Princeton N.J., 287-324.

Aumann, R.J. (1974). "Subjectivity and Correlation in Randomized Strategies", *Journal of Mathematical Economics* **1**, 67-96.

Aumann, R.J. (1987a). "What is Game Theory Trying to Accomplish?", in: K.Arrow and S. Honkapohja (eds.), *Frontiers of Economics*, 28-100.

Aumann, R.J. (1987b). "Game Theory", in: J. Eatwell, M. Milgate and P. Newman (eds.), *The New Palgrave Dictionary of Economics*, 460-482.

Aumann, R.J. (1990). "Nash Equilibria are not Self-Enforcing", in: J.J. Gabszewicz, J.-F. Richard and L.A. Wolsey (eds.), *Economic Decision-Making: Games, Econometrics and Optimisation,* 201-206.

Aumann, R.J. (1992). "Irrationality in Game Theory", in: P. Dasgupta et al., *Economic Analyses of Markets and Games*, MIT Press, Cambridge, 214-227.

Aumann, R.J. (1995). "Backward Induction and Common Knowledge of Rationality", *Games and Economic Behavior* **8**, 6-19.

Aumann, R.J. (1998). "On the Centipede Game", *Games and Economic Behavior* **23**, 97-105.

Aumann, R.J., and A. Brandenburger (1995). "Epistemic Conditions for Nash Equilibrium", *Econometrica* **63**, 1161-1180.

Aumann, R.J., Y. Katznelson, R. Radner, R.W. Rosenthal and B. Weiss (1983). "Approximate Purification of Mixed Strategies", *Mathematics of Operations Research* **8**, 327-341.

Aumann R.J., and M. Maschler (1972). "Some Thoughts on the Minimax Principle", *Management Science* **18**, 54-63.

Aumann R.J., and S. Sorin (1989). "Cooperation and Bounded Recall", *Games and Economic Behavior* **1**, 5-39.

Bagwell, K. and G. Ramey (1996). "Capacity, Entry and Forward Induction", *Rand Journal of Economics* **27**, 660-680.

Balkenborg, D. (1992). *The Properties of Persistent Retracts and Related Concepts*, Ph.D. thesis, Department of Economics, University of Bonn.

Balkenborg, D. (1993). "Strictness, Evolutionary Stability and Repeated Games with Common Interests", CARESS, WP 93-20, University of Pennsylvania.

Banks J.S. and J. Sobel (1987). "Equilibrium Selection in Signalling Games", *Econometrica* **55**, 647-663.

Basu, K. (1988). "Strategic Irrationality in Extensive Games", *Mathematical Social Sciences* **15**, 247-260.

Basu, K. (1990). "On the Non-Existence of Rationality Definition for Extensive Games", *International Journal of Game Theory* **19**, 33-44.

Basu K. and J. Weibull (1991). "Strategy Subsets Closed under Rational Behavior", *Economics Letters* **36**, 141-146.

Battigalli, P. (1997). "On Rationalizability in Extensive Games", *Journal of Economic Theory* **74**, 40-61.

Ben-Porath, E. (1993). "Common Belief of Rationality in Perfect Information Games", Mimeo, Tel Aviv University.

Ben-Porath, E., and E. Dekel (1992). "Signalling Future Actions and the Potential for Sacrifice", *Journal of Economic Theory* **57**, 36-51.

Bernheim, B.D. (1984). "Rationalizable Strategic Behavior", *Econometrica* **52**, 1007-1029.

Bernheim, B.D., B. Peleg and M.D. Whinston (1987). "Coalition-Proof Nash Equilibria I: Concepts", *Journal of Economic Theory* **42**, 1-12.

Bernheim, B.D., and D. Ray (1989). "Collective Dynamic Consistency in Repeated Games", *Games and Economic Behavior* **1**, 295-326.

Bernheim, B.D., and M.D. Whinston (1987). "Coalition-Proof Nash Equilibria II: Applications", *Journal of Economic Theory* **42**, 13-29.

Binmore, K. (1987). "Modeling Rational Players I", *Economics and Philosophy* **3**, 179-214.

Binmore, K. (1988). "Modeling Rational Players II", *Economics and Philosophy* **4**, 9-55.

Blume, A. (1994). "Equilibrium Refinements in Sender-Receiver Games", *Journal of Economic Theory* **64**, 66-77.

Blume, A. (1996). "Neighborhood Stability in Sender-Receiver Games", *Games and Economic Behavior* **13**, 2-25.

Blume, L.E., and W.R. Zame (1994). "The Algebraic Geometry of Perfect and Sequential Equilibrium", *Econometrica* **62**, 783-794.

Börgers, T. (1991). "On the Definition of Rationalizability in Extensive Games", DP 91-22, University College London.

Carlsson, H., and E. van Damme (1993a). "Global Games and Equilibrium Selection", *Econometrica* **61**, 989-1018.

Carlsson, H., and E. van Damme (1993b). "Equilibrium Selection in Stag Hunt Games", in: K. Binmore, A. Kirman and P. Tani (eds.), *Frontiers of Game Theory*, MIT Press, Cambridge, 237-254.

Chin, H.H., T. Parthasarathy and T.E.S. Raghavan (1974). "Structure of Equilibria in n-Person Non-Cooperative Games", *International Journal of Game Theory* **3**, 1-19.

Cho, I.K., and D.M. Kreps (1987). "Signalling Games and Stable Equilibria", *Quarterly Journal of Economics* **102**, 179-221.

Cho, I.K., and J. Sobel (1990). "Strategic Stability and Uniqueness in Signaling Games", *Journal of Economic Theory* **50**, 381-413.

Van Damme, E.E.C. (1983). "Refinements of the Nash Equilibrium Concept", *Lecture Notes in Economics and Math. Systems* **219**, Springer Verlag, Berlin.

Van Damme, E.E.C. (1984). "A Relation Between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games", *International Journal of Game Theory* **13**, 1-13.

Van Damme, E.E.C. (1987a). *Stability and Perfection of Nash Equilibria*. Springer Verlag, Berlin. Second edition 1991.

Van Damme, E.E.C.(1987b). "Equilibria in Non-Cooperative Games", in: H.J.M. Peters and O.J. Vrieze (eds.), *Surveys in Game Theory and Related Topics*, CWI Tract 39, Amsterdam, 1-37.

Van Damme, E.E.C. (1988). "The Impossibility of Stable Renegotiation", *Economics Letters* **26**, 321-324.

Van Damme, E.E.C. (1989a). "Stable Equilibria and Forward Induction", *Journal of Economic Theory* **48**, 476-496.

Van Damme, E.E.C. (1989b). "Renegotiation-Proof Equilibria in Repeated Prisoners' Dilemma", *Journal of Economic Theory* **47**, 206-217.

Van Damme, E.E.C. (1990). "On Dominance Solvable Games and Equilibrium Selection Theories", CentER DP 9046, Tilburg University.

Van Damme, E.E.C. (1992). "Refinement of Nash Equilibrium", in: J.J. Laffont (ed.), *Advances in Economic Theory, 6th World Congress, Vol. 1. Econometric Society Monographs No. 20*, Cambridge University Press, 32-75.

Van Damme, E.E.C. (1994). "Evolutionary Game Theory", *European Economic Review* **38**, 847-858.

Van Damme, E.E.C., and W. Güth (1991a). "Equilibrium Selection in the Spence Signalling Game", in: R. Selten (ed.), *Game Equilibrium Models, Vol. 2: Methods, Morals and Markets*, Springer Verlag, 263-288.

Van Damme, E.E.C., and W. Güth (1991b). "Gorby Games: A Game Theoretic Analysis of Disarmament Campaigns and the Defence Efficiency Hypothesis", in: R. Avenhaus, H. Kavkar and M. Rudniaski (eds.), *Defence Decision Making. Analytical Support and Crises Management*, Springer Verlag, Berlin, 215-240.

Van Damme, E.E.C., and S. Hurkens (1996). "Commitment Robust Equilibria and Endogenous Timing", *Games and Economic Behavior* **15**, 290-311.

Dasgupta, P., and E. Maskin (1986). "The Existence of Equilibria in Discontinuous Games, 1: Theory", *Review of Economic Studies* **53**, 1-27.

Debreu, G. (1970). "Economics with a finite set of equilibria", *Econometrica* **38**, 387-392.

Dierker, E. (1972). "Two Remarks on the Number of Equilibria of an Economy", *Econometrica* **40**, 951-953.

Dold, A. (1972). *Lectures on Algebraic Topology*, Springer Verlag, New York.

Dresher, M. (1961). *Games of Strategy*, Prentice-Hall, Englewood Cliffs, NJ.

Dresher, M. (1970). "Probability of a Pure Equilibrium Point in n-Person Games", *Journal of Combinatorial Theory* **8**, 134-145.

Dubey, P. (1986). "Inefficiency of Nash Equilibria", *Mathematics of Operations Research* **11**, 1-8.

Ellison, G. (1993). "Learning, Local Interaction, and Coordination", *Econometrica* **61**, 1047-1072.

Farrell, J., and M. Maskin (1989). "Renegotiation in Repeated Games", *Games and Economic Behavior* **1**, 327-360.

Forges, F. (1990). "Universal Mechanisms", *Econometrica* **58**, 1341-1364.

Fudenberg, D., D. Kreps and D.K. Levine (1988). "On the Robustness of Equilibrium Refinements", *Journal of Economic Theory* **44**, 354-380.

Fudenberg, D., and D.K. Levine (1993a). "Self-Confirming Equilibrium", *Econometrica* **60**, 523-545.

Fudenberg, D., and D.K. Levine (1993b). "Steady State Learning and Nash Equilibrium", *Econometrica* **60**, 547-573.

Fudenberg, D. and D.K. Levine (1998). *The Theory of Learning in Games*, MIT Press, Cambridge, MA.

Fudenberg, D., and J. Tirole (1991). "Perfect Bayesian Equilibrium and Sequential Equilibrium", *Journal of Economic Theory* **53**, 236-260.

Glazer, J., and A. Weiss (1990). "Pricing and Coordination: Strategically Stable Equilibrium", *Games and Economic Behavior* **2**, 118-128.

Glicksberg, I.L. (1952). "A Further Generalization of the Kakutani Fixed Point Theorem with Application to Nash Equilibrium Points", *Proceedings of the National Academy of Sciences* **38**, 170-174.

Govindan, S. (1995). "Stability and the Chain Store Paradox", *Journal of Economic Theory* **66,** 536-547.

Govindan, S. and R. Robson (1998). "Forward Induction, Public Randomization and Admissibility", *Journal of Economic Theory* **82**, 451-457.

Govindan, S., and R. Wilson (1997). "Equivalence and Invariance of the Index and Degree of Nash Equilibria, *Games and Economic Behavior* **21**, 56-61.

Govindan, S., and R.B. Wilson (1999). "Maximal Stable Sets of Two-Player Games", Mimeo, University of Western Ontario and Stanford University.

Govindan, S., and R. Wilson (2000). "Uniqueness of the Index for Nash Equilibria of Finite Games. Mimeo, University of Western Ontario and Stanford University.

Gul. F., and D. Pearce (1996). "Forward Induction and Public Randomization", *Journal of Economic Theory* **70**, 43-64.

Gul F., D. Pearce and E. Stachetti (1993). "A Bound on the Proportion of Pure Strategy Equilibria in Generic Games", *Mathematics of Operations Research* **18**, 548-552.

Güth, W. (1985). "A Remark on the Harsanyi-Selten Theory of Equilibrium Selection", *International Journal of Game Theory* **14**, 31-39.

Güth, W. (1992). "Equilibrium Selection by Unilateral Deviation Stability", in: R. Selten (ed.), *Rational Interaction, Essays in Honor of John C. Harsanyi*, Springer Verlag, Berlin, 161-189.

Güth, W., and B. Kalkofen (1989). "Unique Solutions for Strategic Games", *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin.

Güth, W., and R. Selten (1991). "Original or Fake - A Bargaining Game with Incomplete Information", in: R. Selten (ed.), *Game Equilibrium Models, Vol. 3: Strategic Bargaining*, Springer Verlag, Berlin, 186-229.

Hammerstein, P., and R. Selten (1994). "Game Theory and Evolutionary Biology", chapter 28 in: R.J. Aumann and S. Hart, (eds.), *Handbook of Game Theory, Vol. 2*, Elsevier, 929-994.

Harris, C. (1985). "Existence and Characterization of Perfect Equilibrium in Games of Perfect Information", *Econometrica* **53**, 613-628.

Harris, C., P. Reny and A. Robson (1995). "The Existence of Subgame-Perfect Equilibrium in Continuous Games with Almost Perfect Information: A Case for Public Randomization", *Econometrica* **63**, 507-544.

Harsanyi, J.C. (1973a). "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed Strategy Equilibrium Points", *International Journal of Game Theory* **2**, 1-23.

Harsanyi, J.C. (1973b). "Oddness of the Number of Equilibrium Points: A New Proof", *International Journal of Game Theory* **2**, 235-250.

Harsanyi, J.C. (1975). "The Tracing Procedure: A Bayesian Approach to Defining a Solution for n-Person Noncooperative Games", *International Journal of Game Theory* **4**, 61-94.

Harsanyi, J.C. (1980). "Analysis of a Family of Two-Person Bargaining Games with Incomplete Information", *International Journal of Game Theory* **9**, 65-89.

Harsanyi, J.C. (1982). "Solutions for Some Bargaining Games under Harsanyi-Selten Solution Theory, Part I: Theoretical Preliminaries; Part II: Analysis of Specific Bargaining Games", *Mathematical Social Sciences* **3**, 179-191, 259-279.

Harsanyi, J.C. (1995). "A New Theory of Equilibrium Selection for Games with Complete Information", *Games and Economic Behavior* **8**, 91-122.

Harsanyi, J.C., and R. Selten (1972). "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information", *Management Science* **18**, 80-106.

Harsanyi, J.C., and R. Selten (1977). "Simple and Iterated Limits of Algebraic Functions", WP CP-370, Center for Research in Management, University of California, Berkeley.

Harsanyi, J.C., and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge, MA.

Hart, S. (1992). "Games in Extensive and Strategic Forms", chapter 2 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 1*, Elsevier, 19-40.

Hart, S. (1999). "Evolutionary Dynamics and Backward Induction", DP 195, Center for Rationality, Hebrew University.

Hart, S., and D. Schmeidler (1989). "Existence of Correlated Equilibria", *Mathematics of Operations Research* **14**, 18-25.

Hauk, E., and S. Hurkens (1999). "On forward induction and evolutionary and strategic stability", WP 408, University of Pompeu Fabra, Barcelona, Spain.

Hellwig, M., W. Leininger, P. Reny and A. Robson (1990). "Subgame-Perfect Equilibrium in Continuous Games of Perfect Information: An Elementary Approach to Existence and Approximation by Discrete Games", *Journal of Economic Theory* **52**, 406-422.

Herings, P.J.J. (2000). "Two Simple Proofs of the Feasibility of the Linear Tracing Procedure", *Economic Theory* **15**, 485-490.

Hillas, J. (1990). "On the Definition of the Strategic Stability of Equilibria", *Econometrica* **58**, 1365-1390.

Hillas, J. (1996). "On the Relation between Perfect Equilibria in Extensive Form Games and Proper Equilibria in Normal Form Games", Mimeo, SUNY, Stony Brook.

Hillas, J. and Kohlberg, E. (1994). "Conceptual Foundations of Strategic Equilibrium", chapter 42 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 3*, Elsevier.

Hillas, J., J. Potters and D. Vermeulen (1999). "On the Relations among some Definitions of Strategic Stability", Mimeo, Maastricht University.

Hillas, J., D. Vermeulen and M. Jansen (1997). "On the Finiteness of Stable Sets: Note", *International Journal of Game Theory* **26**, 275-278.

Hurkens, S. (1994). "Learning by Forgetful Players", *Games and Economic Behavior* **11**, 304-329.

Hurkens, S. (1996). "Multi-Sided Pre-Play Communication by Burning Money", *Journal of Economic Theory* **69**, 186-197.

Jansen, M.J.M. (1981). "Maximal Nash Subsets for Bimatrix Games", *Naval Research Logistics Quarterly* **28**, 147-152.

Jansen, M.J.M., P. Jurg and P. Borm (1990). "On the Finiteness of Stable Sets", Mimeo, University of Nijmegen.

Kalai, E., and D. Samet (1984). "Persistent Equilibria", *International Journal of Game Theory* **13**, 129-141.

Kalai, E., and D. Samet (1985). "Unanimity Games and Pareto Optimality", *International Journal of Game Theory* **14**, 41-50.

Kandori, M., G.J. Mailath and R. Rob (1993). "Learning, Mutation and Long-Run Equilibria in Games", *Econometrica* **61**, 29-56.

Kohlberg, E. (1981). "Some Problems with the Concept of Perfect Equilibrium", NBER Conference on Theory of General Economic Equilibrium, University of California, Berkeley.

Kohlberg, E. (1989). "Refinement of Nash Equilibrium: The Main Ideas", Mimeo, Harvard University.

Kohlberg, E., and J.-F. Mertens (1986). "On the Strategic Stability of Equilibria", *Econometrica* **54**, 1003-1037.

Kohlberg, E., and P. Reny (1997). "Independence on Relative Probability Spaces and Consistent Assessments in Game Trees", *Journal of Economic Theory* **75**, 280-313.

Kreps, D., and G. Ramey (1987). "Structural Consistency, Consistency, and Sequential Rationality", *Econometrica* **55**, 1331-1348.

Kreps, D., and J. Sobel (1994). "Signalling", chapter 25 in: R.J. Aumann and S. Hart, (eds.), *Handbook of Game Theory, Vol. 2*, Elsevier, 849-868.

Kreps, D., and R. Wilson (1982a). "Sequential Equilibria", *Econometrica* **50**, 863-894.

Kreps, D., and R. Wilson (1982b). "Reputation and Imperfect Information", *Journal of Economic Theory* **27**, 253-279.

Kuhn, H.W. (1953). "Extensive Games and the Problem of Information", *Annals of Mathematics Studies* **48**, 193-216.

Lemke, C.E., and J.T. Howson (1964). "Equilibrium Points of Bimatrix Games", *Journal of the Society for Industrial and Applied Mathematics* **12**, 413-423.

Leopold-Wildburger, U. (1985). "Equilibrium Selection in a Bargaining Problem with Transaction Costs", *International Journal of Game Theory* **14**, 151-172.

Madrigal, V., T. Tan and S. Werlang (1987). "Support Restrictions and Sequential Equilibria", *Journal of Economic Theory*, **43**, 329-334.

Mailath, G.J., L. Samuelson and J.M. Swinkels (1993). "Extensive Form Reasoning in Normal Form Games ", *Econometrica* **61**, 273-302.

Mailath, G.J., L. Samuelson and J.M. Swinkels (1997). "How Proper is Sequential Equilibrium", *Games and Economic Behavior* **18**, 193-218.

Maynard Smith, J., and G. Price (1973). "The Logic of Animal Conflict", *Nature* **246**, 15-18.

McLennan, A. (1985). "Justifiable Beliefs in Sequential Equilibrium", *Econometrica* **53**, 889-904.

Mertens, J.-F. (1987). "Ordinality in Non-Cooperative Games", CORE DP 8728, Université Catholique de Louvain, Louvain-la-Neuve.

Mertens, J.-F. (1989a). "Stable Equilibria - A Reformulation, Part I, Definition and Basic Properties", *Mathematics of Operations Research* **14**, 575-625.

Mertens, J.-F. (1989b). "Equilibrium and Rationality: Context and History-Dependence", CORE DP, October 1989, Université Catholique de Louvain, Louvain-la-Neuve.

Mertens, J.-F. (1990). "The 'Small Worlds' Axiom for Stable Equilibria", CORE DP 9007, Université Catholique de Louvain, Louvain-la-Neuve.

Mertens, J.-F. (1991). "Stable Equilibria - A Reformulation, Part II, Discussion of the Definition, and Further Results", *Mathematics of Operations Research* **16**, 694-753.

Mertens, J.-F. (1992). "Two Examples of Strategic Equilibrium", CORE DP 9208, Université Catholique de Louvain, Louvain-la-Neuve.

Milgrom, P., and D.J. Roberts (1982). "Predation, Reputation and Entry Deterrence", *Journal of Economic Theory* **27**, 280-312.

Milgrom, P., and D.J. Roberts (1990). "Rationalizability, Learning and Equilibrium in Games with Strategic Complementarities", *Econometrica* **58**, 1255-1278.

Milgrom, P., and D.J. Roberts (1991). "Adaptive and Sophisticated Learning in Repeated Normal Form Games", *Games and Economic Behavior* **3**, 1255-1278.

Milgrom, P., and C. Shannon (1994). "Monotone Comparative Statics", *Econometrica* **62**, 157-180.

Milgrom, P., and R. Weber (1985). "Distributional Strategies for Games with Incomplete Information", *Mathematics of Operations Research* **10**, 619-632.

Monderer, D., and L.S. Shapley (1996). "Potential Games", *Games and Economic Behavior* **14**, 124-143.

Moulin, H. (1979). "Dominance Solvable Voting Games", *Econometrica* **47**, 1337-1351.

Moulin, H., and J.P. Vial (1978). "Strategically Zero-Sum Games: The Class of Games Whose Completely Mixed Equilibria Cannot Be Improved upon", *International Journal of Game Theory* **7**, 201-221.

Myerson, R. (1978). "Refinements of the Nash Equilibrium Concept", *International Journal of Game Theory* **7**, 73-80.

Myerson, R.B. (1986). "Multistage Games with Communication", *Econometrica* **54**, 323-358.

Myerson, R.B. (1994). "Communication, Correlated Equilibria, and Incentive Compatibility", chapter 24 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 2*, Elsevier, 827-848.

Nash, J.F. (1950a). *Non-Cooperative Games*, Ph.D. Dissertation, Princeton University.

Nash, J.F. (1950b). "Equilibrium Points in n-Person Games", *Proceedings from the National Academy of Sciences, U.S.A.*, **36**, 48-49.

Nash, J.F. (1951). "Non-Cooperative Games", *Annals of Mathematics* **54**, 286-295.

Nash, J.F. (1953). "Two-Person Cooperative Games", *Econometrica* **21**, 128-140.

Neumann, J. von, and O. Morgenstern (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton NJ (first edition 1944).

Neyman, A. (1997). "Correlated Equilibrium and Potential Games", *International Journal of Game Theory* **26**, 223-227.

Noldeke, G., and E.E.C. van Damme (1990). "Switching Away From Probability One Beliefs", University of Bonn DP A-304.

Norde, H. (1999). "Bimatrix Games have Quasi-Strict Equilibria", *Mathematical Programming* **85**, 35-49.

Norde, H., J. Potters, H. Reijnierse and D. Vermeulen (1996). "Equilibrium Selection and Consistency", *Games and Economic Behavior* **12**, 219-225.

Okada, A. (1983). "Robustness of Equilibrium Points in Strategic Games", DP B137, Tokyo Institute of Technology.

Osborne, M. (1990). "Signaling, Forward Induction, and Stability in Finitely Repeated Games", *Journal of Economic Theory* **50**, 22-36.

Pearce, D. (1984). "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* **52**, 1029-1050.

Peleg, B. and S.H. Tijs (1996). "The Consistency Principle for Games in Strategic Form", *International Journal of Game Theory* **25**, 13-34.

Ponssard, J.-P. (1991). "Forward Induction and Sunk Costs Give Average Cost Pricing", *Games and Economic Behavior* **3**, 221-236.

Radner, R., and R.W. Rosenthal (1982). "Private Information and Pure Strategy Equilibria", *Mathematics of Operations Research* **7**, 401-409.

Raghavan, T.E.S. (1994). "Zero-Sum Two-Person Games", chapter 20 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 2*, Elsevier, 735-768.

Reny, P. (1992a). "Backward Induction, Normal Form Perfection and Explicable Equilibria", *Econometrica* **60**, 627-649.

Reny, P.J. (1992b). "Rationality in Extensive Form Games", *Journal of Economic Perspectives* **6**, 103-118.

Reny, P.J. (1993). "Common Belief and the Theory of Games with Perfect Information", *Journal of Economic Theory* **59**, 257-274.

Ritzberger, K. (1994). "The Theory of Normal Form Games from the Differentiable Viewpoint", *International Journal of Game Theory* **23**, 207-236.

Rosenmüller, J. (1971). "On a Generalization of the Lemke-Howson Algorithm to Noncooperative n-Person Games", SIAM *Journal of Applied Mathematics* **21**, 73-79.

Rosenthal, R. (1981). "Games of Perfect Information, Predatory Pricing and the Chain Store Paradox", *Journal of Economic Theory* **25**, 92-100.

Rubinstein, A. (1989). "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge' ", *American Economic Review* **79**, 385-391.

Rubinstein, A. (1991). "Comments on the Interpretation of Game Theory", *Econometrica* **59**, 909-924.

Rubinstein, A., and A. Wolinsky (1994). "Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability", *Games and Economic Behavior* **6**, 299-311.

Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*, MIT Press, Cambridge, MA.

Schanuel, S.H., L.K. Simon and W.R. Zame (1991). "The Algebraic Geometry of Games and the Tracing Procedure", in: R. Selten (ed.), *Game Equilibrium Models, Vol. 2: Methods, Morals and Markets*, Springer Verlag, Berlin, 9-43.

Selten, R. (1965). "Spieltheoretische Behandlung eines Oligopolmodels mit Nachfragetragheit", *Zeitschrift für die Gesamte Staatswissenschaft* **12**, 301-324, 667-689.

Selten, R. (1975). "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory* **4**, 25-55.

Selten, R. (1978). "The Chain Store Paradox", *Theory and Decision* **9**, 127-159.

Selten, R. (1995). "An Axiomatic Theory of a Risk Dominance Measure for Bipolar Games with Linear Incentives", *Games and Economic Behavior* **8**, 213-263.

Selten, R., and W. Güth (1982). "Equilibrium Point Selection in a Class of Market Entry Games", in: M. Deistler, E. Fürst and G. Schwödiauer (eds.), *Games, Economic Dynamics, and Time Series Analysis - A Symposium in Memoriam of Oskar Morgenstern*, Physica Verlag, Würzburg, 101-116.

Selten, R., and U. Leopold (1983). "Equilibrium Point Selection in a Bargaining Situation with Opportunity Costs", *Economie Appliqueé* **36**, 611-648.

Shapley, L.S. (1974). "A Note on the Lemke-Howson Algorithm", *Mathematical Programming Study* **1**, 175-189.

Shapley, L.S. (1981). "On the Accessibility of Fixed Points", in: O. Moeschlin and D. Pallaschke (eds.), *Game Theory and Mathematical Economics*, North Holland Publishing Company, Amsterdam, 367-377.

Shubik, M. (2001). "Game Theory and Experimental Gaming", chapter 62 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory*, Vol. 3, Elsevier.

Simon, L.K., and M.B. Stinchcombe (1995). "Equilibrium Refinement for Infinite Normal-Form Games", *Econometrica* **63**, 1421-1443.

Simon, L.K. and W.R. Zame (1990). "Discontinuous Games and Endogenous Sharing Rules", *Econometrica* **58**, 861-872.

Sorin, S. (1992). "Repeated Games with Complete Information", chapter 4 in: R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory, Vol. 1*, Elsevier, 71-108.

Spence, M. (1973). "Job Market Signalling", *Quarterly Journal of Economics* **87**, 355-374.

Stanford, W. (1995). "A Note on the Probability of k Pure Nash Equilibria in Matrix Games", *Games and Economic Behavior* **9**, 238-246.

Tarski, A. (1955). "A Lattice Theoretical Fixed Point Theorem and its Applications", *Pacific Journal of Mathematics* **5**, 285-308.

Topkis, D. (1979). "Equilibrium Points in Nonzero-Sum n-Person Submodular Games", SIAM *Journal of Control and Optimization* **17**, 773-787.

Vega-Redondo, F. (1996). *Evolution, Games and Economic Behavior*, Oxford University Press, Oxford, UK.

Vives, X. (1990). "Nash Equilibrium with Strategic Complementarities", *Journal of Mathematical Economics* **19**, 305-321.

Weibull, J. (1995). *Evolutionary Game Theory*, MIT Press, Cambridge, MA.

Wilson, C. (1977). "A Model of Insurance Markets with Incomplete Information", *Journal of Economic Theory* **16**, 167-207.

Wilson, R.B. (1971). "Computing Equilibria of n-Person Games", SIAM *Journal of Applied Mathematics* **21**, 80-87.

Wilson, R.B. (1992). "Computing Simply Stable Equilibria", *Econometrica* **60**, 1039-1070.

Wilson, R.B. (1997). "Admissibility and Stability", in: W. Albers et al., *Understanding Strategic Interaction; Essays in Honor of Reinhard Selten*, Springer Verlag, Berlin, 85-99.

Young, P. (1993a). "The Evolution of Conventions", *Econometrica* **61**, 57-84.

Young, P. (1993b). "An Evolutionary Model of Bargaining", *Journal of Economic Theory* **59**, 145-168.

Zermelo, E. (1912). "Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels", in: E.W. Hobson and A.E.H. Love (eds.), *Proceedings of the Fifth International Congress of Mathematicians, Vol. 2*, Cambridge University Press, 501-504.